NETWORK MODELS FOR CAPTURING MOLECULAR FEATURE AND PREDICTING DRUG TARGET FOR VARIOUS CANCERS

Enze Liu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

December 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Xiaowen Liu, PhD, Co-Chair

_____

Huanmei Wu, PhD, Co-Chair

_____

Chi Zhang, PhD

September 25, 2020

_____

Jun Wan, PhD

_____

Sha Cao, PhD

_____

Lang Li, PhD

ACKNOWLEDGEMENTS

I would like to express my most sincere appreciation to my advisor Professor Lang Li. Without his guidance, none of these studies can be conducted, not to mention the PhD thesis work. His great mentorship and rigorous attitude towards scientific research have forged me. I also would like to send my gratitude to my committee members: Professor Huanmei Wu, Professor Xiaowen Liu, Professor Chi Zhang, Professor Jun Wan and Professor Sha Cao. Their comments and suggestions are priceless. I also would like to thank the School of Informatics and Computing in Indiana University for their tremendous effect on designing the courses and providing assistance for all students.

I also would like to thank my colleagues, Dr. Lei Wang, Dr. Xue Wu, Dr. Pengyue Zhang, Dr. Hen-yi Wu, Mr. Yang Huo, Mr. Chuanpeng Dong. It has been very educating and fun to collaborate with you.

My appreciation also goes to my parents: Guoming Liu and Jiajia Li, and all my relatives and friends, for their undoubtable support. Last but not the least, I would like to thank my fiancée Shijun Zhang. Thank you for being by my side.

Enze Liu

NETWORK MODELS FOR CAPTURING MOLECULAR FEATURE AND PREDICTING DRUG TARGET FOR VARIOUS CANCERS

Network-based modeling and analysis have been widely used for capturing molecular trajectories of cellular processes. For complex diseases like cancers, if we can utilize network models to capture adequate features, we can gain a better insight of the mechanism of cancers, which will further facilitate the identification of molecular vulnerabilities and the development targeted therapy. Based on this rationale, we conducted the following four studies:

A novel algorithm ‘FFBN’ is developed for reconstructing directional regulatory networks (DEGs) from tissue expression data to identify molecular features. ‘FFBN’ shows unique capability of fast and accurately reconstructing genome-wide DEGs compared to existing methods. FFBN is further used to capture molecular features among liver metastasis, primary liver cancers and primary colon cancers. Comparisons among these features lead to new understandings of how liver metastasis is similar to its primary and distant cancers.

‘SCN’ is a novel algorithm that incorporates multiple types of omics data to reconstruct functional networks for not only revealing molecular vulnerabilities but also predicting drug targets on top of that. The molecular vulnerabilities are discovered via tissue-specific networks and drug targets are predicted via cell-line specific networks. SCN is tested on primary pancreatic cancers and the predictions coincide with current treatment plans.

‘SCN website’ is a web application of ‘SCN’ algorithm. It allows users to easily submit their own data and get predictions online. Meanwhile the predictions are displayed along with network graphs and survival curves.

‘DSCN’ is a novel algorithm derived from ‘SCN’. Instead of predicting single targets like ‘SCN’, ‘DSCN’ applies a novel approach for predicting target combinations using multiple omics data and network models.

In conclusion, our studies revealed how genes regulate each other in the form of networks and how these networks can be used for unveiling cancer-related biological processes. Our algorithms and website facilitate capturing molecular features for cancers and predicting novel drug targets.

Xiaowen Liu, PhD, Co-Chair

Huanmei Wu, PhD, Co-Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Modern system biology focuses on understanding how genes and other molecules work in concert as a complex system to form and regulate biological processes in every living organism [1]. These regulations consist of various types: Cis-regulatory elements and trans-regulatory elements are two main regulatory types [2]. Cis-regulatory elements are present near the structural portion of the gene/protein as the gene they regulate, such as the photosynthetic protein family, are expressed at the same time in development. Whereas trans-regulatory elements can distantly regulate genes from which they were transcribed. Enhancers and multiple trans-acting factors are essential for trans-control transcription initiations. Regulators contain DNA epigenetic modifications by methylation, miRNA, transcription factors (TFs), and post-translational modification (PTM) that include histone proteins and other proteins, which are involved in methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation (Figure 1). A motif is a sequence pattern that carries out certain functions in DNA, RNA and proteins. Normally, TFs coded by a gene can regulate gene expression by binding to specific motifs. MicroRNAs (miRNAs) regulate gene expression via RNA silencing or post-transcription regulations. Histones can alter the chromatin structure, which further controls the access of TFs and polymerases to genes thus resulting in an expression regulation [3]. Methylation plays a crucial role in regulating gene expression by blocking the promoters that can activate TFs [4]. Large experimental evidence has been gathered to verify biology gene interactions, such as

KEGG [5], Pathway Commons [6], MetaCyc Metabolic Pathway Database [7], JASPAR CORE [8], HistoneDB 2.0 [9], and miRGator v3.0 [10], GeneHancer [11], etc.

A GRN is composed of functional linkages between regulators and targets on genome, transcriptome, and proteome levels (Figure 1). All stated interaction types can be integrated into GRNs to be studied systematically. Thus, Accurate reconstructing a GRN is one of the key tasks in systems biology. Recent advances in high-throughput techniques provide an opportunity to reconstruct regulatory networks, by offering a huge amount of binding data, like DNA-Seq, RNA-Seq, ChIP-Seq, and Mass spectrometry data [12]. Subsequently, various studies have been conducted for reconstructing GRNs using single or combinations of these types of data. For instance, Tong et. al. utilizes experiment-derived results and computational predictions to reconstruct a genetic interaction network that covers the entire yeast genome [13]. Margolin et. al. reconstructed a genome-wide co-expression network for humans by applying mutual information theory [14]. Stelzl et. al. reconstructed the whole proteome protein-protein interaction network using yeast two-hybrid system [15]. However, GRNs reconstructed by most of the previous work are general GRNs for different organisms, which cannot completely reflect the cellular machinery under the influence of diseases, i.e. cancers.

To better understand the cellular processes in a context-specific setting, in this thesis, we attempt to reconstruct cancer-specific regulatory networks from various omics-data to reveal cancer-related features from a network perspective. In section 2, different types of GRNs and the underlying theory for predicting them are introduced.

Evaluation schemes and applications of these GRN models are described. In section 3, an approach called ‘FFBN’ is introduced for reconstructing GRNs using Bayesian network models for identifying molecular similarities and differences between primary cancers and metastatic cancers using gene expression data. In section 4, ‘SCN’ approach is introduced as a tool for reconstructing functional networks for predicting drug targets for cancers using omics-data from all three levels of the central dogma. In section 5, a web application ‘SCNwebsite’ is introduced as a web extension of ‘SCN’ that allows users to conveniently predict drug targets online. In section 6, an algorithm called ‘DSCN’, which is the extension work of ‘SCN’, for predicting target combinations for cancers, is introduced.

2 Background

2.1 Reconstructing Gene regulatory networks

Reconstructing a GRN that incorporates genes and other molecules help to gain a better understanding of biological processes underlying every living organism [16]. Currently, various studies have been conducted for reconstructing GRNs at each level of central dogma (Figure 2). For instance, Baryshnikova et al utilized DNA-seq data to generate genetic interaction network for yeast to better understanding genetic variants responsible for disease in humans [17]. Madan et al reconstructed co-expression networks to systematically study interactions between transcription factors and their target genes in E.coli from expression data [18]. Databases like STRING [19] offer Protein-protein interaction networks, in which interactions are curated from experiments and existing databases. However, A functional network, which is usually generated from the integration of multiple types of omics data, can precisely capture the heterogeneity of diseases and biological processes from a comprehensive perspective. Wang et al. integrated expression and methylation data to generate a functional network that reveals subtypes of cancer, which hasn't been classified before [20]. are The inference of a GRN is often accomplished through the use of gene expression data. So far, there are numerous computational methods and models developed for restoring GRNs in a real cellular environment. However, each of them has their assumptions and methods, drawing different blueprints that the GRN described. There is still much confusion about the basic meaning of GRN, ways of assessment, and possible biomedical application. Typically, the relationships between

genes are directional in nature and they can change over time or in response to external stimulus. Researchers are facing the choice of whether to include extra features such as causality and temporal behaviors when modeling gene networks or not.

2.1.1 Gene interaction network

Genes are said to be interacting if mutations in two or more genes cause phenotypical changes [21]. Positive gene interactions indicate beneficial events such as two mutations generating less severe phenotype while negative gene interactions suggest defective events such as synthetic lethality or reduced fitness [22]. Constructing gene interaction networks can illustrate how consistent sets of negative or positive gene interactions are when connecting protein complex and pathways to map a functional wiring diagram [21]. Previously, screening techniques can only analyze interactions within a relatively simple genome such as yeast. Recently, genome-wide application of CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 screening approaches have enabled researchers to draw whole-genome genetic interaction networks in human cells [23-25].

2.1.2 Co-expression network

A GRN built with a correlation-based method from RNA data is called a co-expression network [26]. A co-expression network is constructed by calculating pairwise correlations for each gene pairs to form a fully connected graph, and a threshold scheme is applied to remove edges between genes that are not significantly correlated (Figure. 3a). Pearson Correlation, Spearman Correlation, and Euclidean

Distance are widely used for measuring linear correlations while Mutual Information (MI) and kernel correlations between two genes are the representative measures for detecting non-linear correlations [27]. P-value, Z-score, clustering coefficient [28], and random matrix theory [29] are typical threshold schemes for filtering out low content co-expression.

Pearson correlation ρ is the most basic correlation that measures the pairwise correlation between two continuous variables.

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

Here X and Y denote expression levels of two genes. Cov is the covariance between gene X, Y, and sigma σ indicates the individual standard deviation. Pearson correlation measures the collinearity between two genes with the assumption that both genes' expression levels follow an approximately normal distribution, making it unsuitable for measuring genes with nonlinear relationships.

Rank-based correlation describes correlations by comparing the rank of the variables (genes) instead of covariance. Take Spearman Correlation as an example, it compares the monotonic relationship between two variables (gene). If two genes follow the normal expression pattern and have a clear linear relationship, then Pearson Correlation and Spearman Correlation would be very similar. However, if two genes are monotonically correlated not linearly correlated, only Spearman Correlation could measure the non-linear relationships. In other words, Spearman Correlation is more general and robust compared to the Pearson Correlation, even though information gets lost during the ranking transformation.

Distance-based correlation mainly refers to Euclidean Distance, which is the geometric distance between two genes if two genes are assumed to be in the same p-dimensional Euclidean space. Euclidean distance is sensitive to transformations, whereas the other two are normally invariant to linear transformations.

Information-theoretic entropy and mutual information concepts provide a new perspective to scale the correlation between variables. Entropy measures the quantity of information within a given system and two variables that contain more entropy tend to be more related.

Mutual information (MI) measures all the entropy within a system, including marginal entropy from variables and joint entropy between/among variables. A MI $I(X,Y)$ measures a system of two variables (genes) and can be defined as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \ (2)$$

where

$$H(X) = -\sum_{x=1}^{m} f(x)\log\big(f(x)\big) \ (3)$$

is the marginal entropy and:

$$H(X,Y) = -\sum_{x=1,y=1}^{m.n} f(x,y)\log\big(f(x,y)\big) \ (4)$$

is the joint entropy for two continuous variables of gene X and Y.

Relevance network A gene association network built with MI is called a relevance network (Butte and Kohane, 2000). ARACNe (Margolin et al., 2006) is the most famous and representative algorithms for building a relevance network. ARACNe assumes genes following a normal distribution and performed a Gaussian Kernel

estimator (Scott and Sheather, 1985) to estimate marginal distribution f(x) and joint distribution f(x,y) in formula 3 and 4. ARACNe defined its edge-pruning thresholds named data processing inequality (DPI) as: for all available triplets $< i, j, k >$, if i interacts with k via j, then the edges (pairwise MI) among them should satisfy the following inequality:

$$I(i,j) \leq min(I(i,k), I(k,j)) \quad (5)$$

meaning that the indirect interaction in the system should contain the smallest MI value. Any violated indirect interactions will be pruned.

GRN based on correlations provide co-expression or co-functionality from a network scale with no directionality, meaning that if no prior knowledge is provided (e.g. a gene is confirmed to be a transcription factor in a two-gene interaction system). In this case, causations among genes in a GRN cannot be reflected. MI detects underlying correlations that linear Pearson Correlation cannot discover, especially for variables with non-linear correlation.

2.1.3 Protein-protein interaction network

Protein-protein interaction (PPI) network reflects a network representation of all PPIs within a system, e.g. an organism's cell. PPIs are either experimentally verified or inferred by their properties. Experiments such as yeast two-hybrid screening [30] and affinity purification couple to mass spectrometry [31] can detect *in vivo* PPIs. Inference based methods predict PPIs based on their structure [32], functional domain [33], or binding sites information [34]. The interactions (edges) in PPI networks can be

weighted and directed. The weights indicate the belief in an association [35] while directions provide information about the signal flow [36].

2.2. Network models

Figure 3 lists some representative network models in current day's GRN reconstruction. In this figure, a column indicates a co-expression network; b column indicates a regression-based regulatory network; c column indicates a Bayesian network model; d column indicates a differential equation-based regulatory network; e column indicates dynamic Bayesian network

2.2.1 Bayesian network

A Bayesian network (BN) is a probabilistic network model that takes a group of random variables and measures their relationships as a directed acyclic graph (DAG) with all the conditional dependencies among them (Figure. 3(c)). Given a certain network structure G, a BN can be expressed as the following joint probability density function (PDF):

$$p_B(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i | P_{a_i}^{g}, \theta_i) \quad (6)$$

Where $P_{a_i}^{g}$ is the set of parents of gene node x_i in a graph G, node $g \in G$, θ_i is the i-th distribution mean, where all gene expression obeys a normal distribution N(θ_i, δ_i). The product of them indicates the local Markov property of BN: Every node (gene) is conditionally independent of its non-descendants given its parents. Reconstructing a BN from the given expression data requires two types of inferences: (1) Parameter learning, which infers marginal and conditional distributions of all nodes, and (2)

structure learning, which infers the optimal topology that has the biggest overall probability.

Parameter learning In parameter learning, the marginal distributions can be either given by prior knowledge or inferred by various methods such as the principle of maximum entropy [37] while conditional distributions are often inferred with approaches such as maximum likelihood estimator or expectation maximization [38].

Structure learning Learning the optimal structure of a BN has been a great challenge for decades. One can always use a brute force approach to go through every possible structure and finally identify the best, which would result in exponential time complexity to achieve. Thus, most of the existing algorithms for searching the optimal structures are heuristic and progressive. For instance, Monte Carlo Markov chain (MCMC) methods use approximate sampling approaches [39]. Greedy equivalence search (GES) proposed a local optimal constraint and performed a stepwise searching scheme [40].

Advantages and bottlenecks Compared to undirected co-expression networks, BNs can point out the directionality of each edge thus revealing a causation relationship among all genes, which is significantly advantageous. Large number of experiments have shown that BN can offer better accuracy and tolerance with noise expression input on a small number of genes [41]. As more variables (genes) are incorporated into the model, the computational time grows exponentially, which brings up the great challenge of building a whole genomic BN for higher organisms like humans [41]. Moreover, for one dataset, there might be more than one optimal BN

structure that has the equivalent overall probabilities. These models (structures) are called equivalence class [40], in which each model is probabilistically indistinguishable. For instance, two BNs $A \rightarrow B \rightarrow C$ and $C \rightarrow B \rightarrow A$ are probabilistically equivalent since their overall probabilities are $P(A)P(B|A)P(C|B)$ and $P(C)P(B|C)P(A|B)$ are the same. However, their directions are completely reversed.

Causal network Causal network (CN) refers to a stricter form of BN, meaning not only conditional dependencies but also Markov conditions among variables need to be followed in the DAG, which provide clear causation of one variable over another. Hence, the above BN are not causally equivalent due to different Markov conditions: $A\_||\_B|C$ (A is independent of C given B) versus $C\_||\_A|B$ (C is independent of A given B), where in the first network, A causes C, in the second, C causes A. Hence, causal network inference provides a more explicit interpretation of relationships among variables compared to BN.

D-separation D-separation measures a state of dependencies of two variables in a triplet. In the following connected triplets (7), (8), (9), and (10), (10) is defined D-separated while the rest are called D-connected. According to Pearl [42], in (4), B blocks the information transmission between A and C.Thus in the case, A and C are independent and Vice versa. Given this property, for a triplet with undirected edges A -- B-- C, if the calculated correlation between A and C is lower than a given threshold, then the edges can be quickly oriented as $A \leftarrow B \rightarrow C$, which saves a great deal of computational time.

$$A \leftarrow B \rightarrow C \quad (7)$$

$$A \leftarrow B \leftarrow C \quad (8)$$

$$A \rightarrow B \rightarrow C \quad (9)$$

$$A \rightarrow B \leftarrow C \quad (10)$$

2.2.2 Functional network

While PPI networks provide evidence on how proteins physically interact with each other, functional networks connect genes with related functions even if they do not physically interact [43]. Functional GRNs are constructed to depict both direct and indirect regulations among genes. These regulations can be based on genetic interactions, co-expression, PPIs, and pathway data. Functional networks built with combinations of these data can provide comprehensive views for given phenotypes or diseases. For instance, co-expression data are frequently in concert with PPIs to obtain functional modules in yeast and human [44-46]. Combining pathways as skeleton and co-expression data as edge weights generate functional networks that reveal regulation differences among different cohorts, which further predict drug/target combinations [47-49].

2.2.3 Network analyses

Degree Degree refers to how many connected neighbors a node has in a network. In a directed network, degree of a node is further divided into in-degree and out-degree. A node having significantly high degrees are considered hub node. In a GRN, a hub node usually provides higher impact towards the network than a non-hub node does. Degree distributions also reflect properties of a network. For instance, many biological networks have scale-free structure, in which very few nodes have very

high degrees while the other nodes have low degrees. A scale-free network has high resistance to network paralyze caused by random node failure. But if failures occur on hub nodes, the network will generally lose connectedness [50].

Module A Network module is defined as a subnetwork of highly connected nodes. Studying module properties and their changes across conditions help to identify similarities and differences between two networks [51]. These similarities/differences may lead to the breakthrough of biological problems. For instance, network modules are generated using multiple omics data and compared for identifying novel subtypes of various cancers [20]. On the other hand, modules (subnetworks) are simpler than the whole network. Studying modules can efficiently reduce the complexity and background noise introduced by the whole network.

2.3 Model evaluation and Validations

Evaluation metrics A Bayesian information criterion (BIC) or Akaike's information criterion (AIC) is chosen for measuring the goodness of fit of the learned structure [52]. BIC is defined as:

$$BIC = nln\left(\frac{RSS}{n}\right) + kln(n) \quad (11)$$

Where RSS is the residual sum of square, n is the sample size and k is the number of parameters (nodes) in the model. To measure the accuracy of the model, Receiver operating characteristics (ROC) curve and Precision-Recall (PR) curve are often used. ROC curve is created by plotting true positive rate (X-axis) against false positive rate (Y-axis) under various threshold settings. Similarly, precision-Recall curve is defined

as plotting precision against recall under various thresholds. The four components of plotting these two curves are defined as follows:

True Positive Rate=True Positives / (True Positives + False Negatives) (12)

False Positive Rate= False Positives / (False Positives + True Negatives) (13)

Precision = True Positives / (True Positives + False Positives) (14)

Recall= True Positives / (True Positives + False Negatives) (15)

Gold standards In computer science, Gold standard refers to criteria that is supported by scientific evidence. Typically, a gold standard contains both positive and negative set. A good algorithm or model should identify as many members as possible from positive set and vice versa from negative set. The quantity and quality of a gold standard can largely impact the prediction accuracy. For GRN reconstruction, the most common gold standards are from validated experiment results such as pathways and PPIs. Databases like Kyoto encyclopedia of genes and genomes (KEGG) [5] and Gene Ontology (GO) [53] provide validated gene interactions in the form of pathways. Databases like BioGrid [54] and STRING [32] offers evidence of PPI and protein complexes. Under given conditions, associations within these databases can be used as positive set and any associations beyond can be used as negative set. Notably, in 'Dialogue on Reverse Engineering Assessment and Methods' (DREAM) network inference challenge [55], validated regulations and regulation intensities are provided between certain transcription factors and target genes. Any regulations beyond them are considered as false regulation. Hence, a positive and a negative set are generated. Algorithms and models are tested with two sets to identify TP, FP, TN and FN. Under

all possible parameters, ROC curves and PR curves can be drawn. And the overall performance of algorithms and models can be measured by Area Under ROC (AUROC) and Area Under PR (AUPR) accordingly.

2.4 Drug targets discovery

Single drug target selection has been extensively studied recently while various methods have been developed. For instance, 'Connectivity map' project (C-map) curated expression profiles of human cells exposed to thousands of drugs, which can be served for drug repositioning [56]. Ma et. al developed an algorithm named 'Met-express' [57] that combines gene co-expression network with human metabolic network to predict drug targets for pancreatic cancer. However, these methods only utilize expression data as fundamental knowledge and incorporate other biological knowledge to predict targets. However, most drugs function on protein level eventually. And expression level regulation might not eventually reflect on protein level. Secondly, their analysis lacks the support of cell survival phenotypes that directly reflect the effect of gene knockdown/knockout experiment.

Combinational target wise, there are two existing studies that infer synergistic effect of combinational proteins that can be potentially treated as combinational targets. For instance, 'OptiCon' algorithm [49] takes advantage of existing pathways and mutation information to generate functional networks from expression data. Functional networks are further divided into subnetworks using bipartite network simplification approaches. Optimal control node for each network is identified and synergistic scores among optimal control nodes are calculated. 'VIPER' algorithm [58] applies the Master

Regulator Inference algorithm for identifying core regulators from a gene regulatory network generated from gene expression data. These synergistic core regulators can potentially be considered as candidate combinational targets. These methods focus on investigating the synergistic effect among protein, which can be treated as potential combinational targets. However, neither methods consider the durability of these protein combinations. Majority of their works are conducted on gene expression level. Other genomic properties like CRISPR-Cas9 screening efficiency of selected combinations are not evaluated or estimated, which might result in a failure of applying these combinations *in vitro* or in clinic.

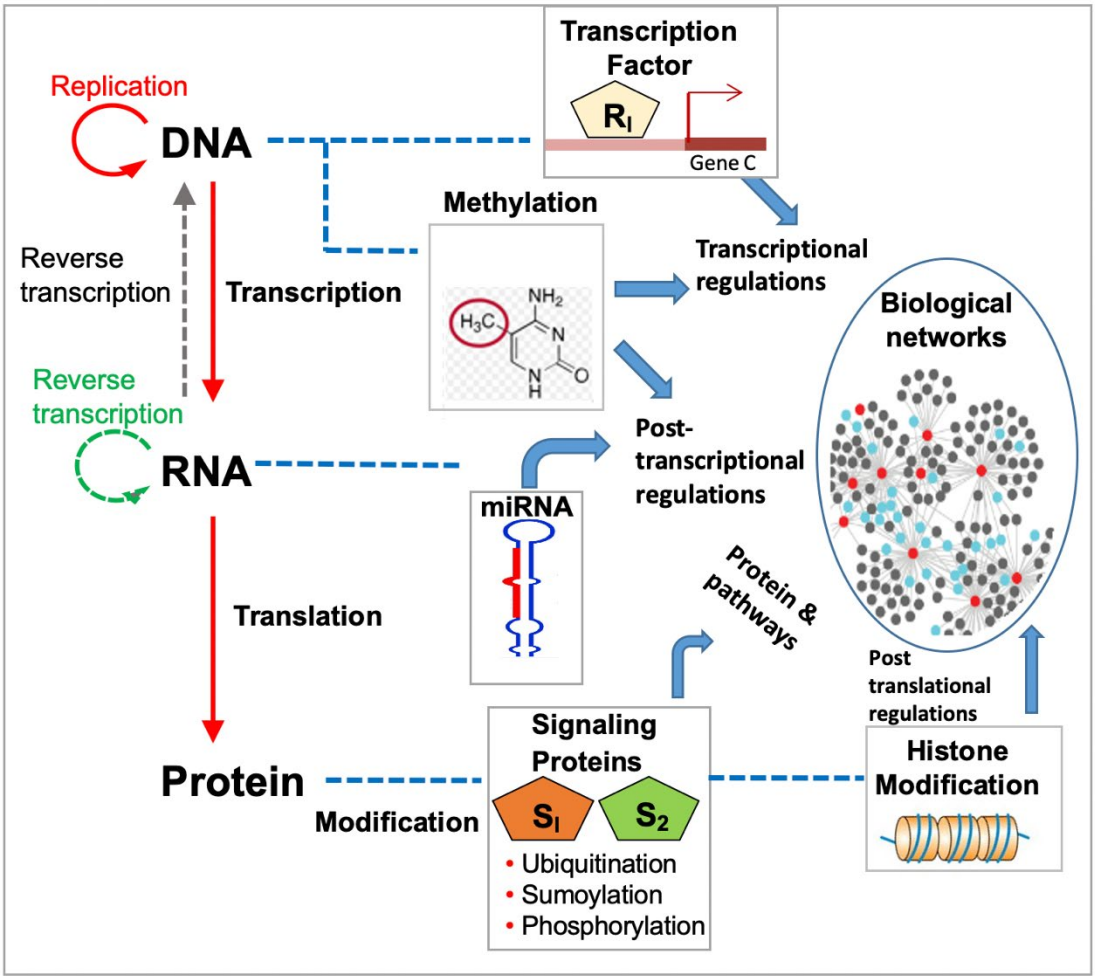

Figure 1 Central dogma and main regulate elements for biological networks

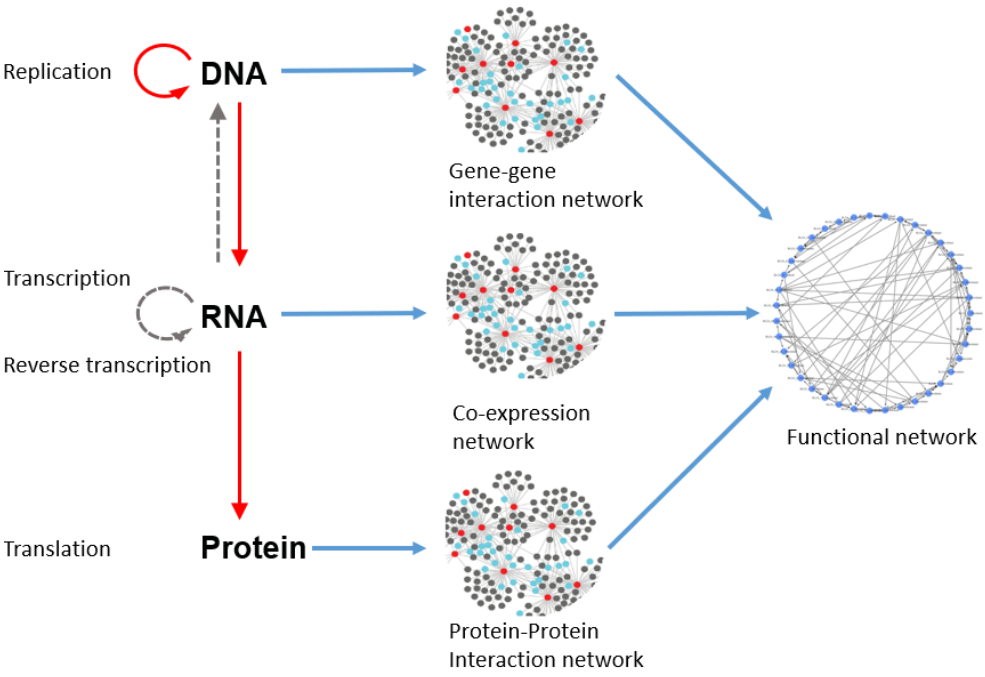

Figure 2 Central dogma and corresponding networks at each level

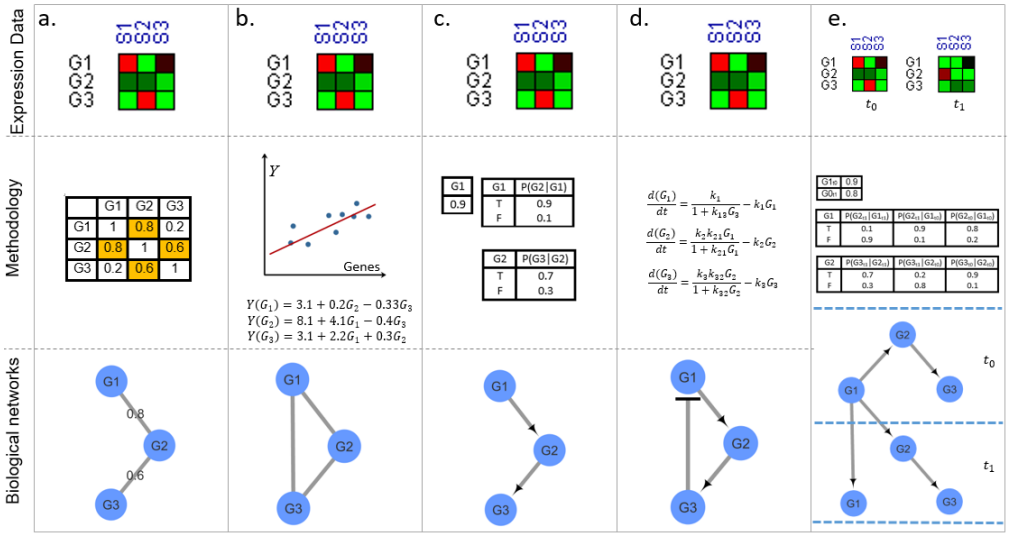

Figure 3 Representative models for reconstructing regulatory networks

3 A fast and Furious Bayesian Network (FFBN) and Its Application to Identify Colon Cancer to Liver Metastasis Molecular Features

3.1 Introduction

Bayesian network (BN) is notoriously known for its slowness in achieving global optimal, especially when it is applied to the transcriptome data where thousands of genes are clustering together. However, BN is a powerful tool to discover regulatory relationships, especially suitable for biological conditions where the molecular mechanisms are largely unknown. One example is the colon to liver metastasis. Biologically, it is still unknown whether colon to liver metastasis is colon cancer or liver cancer. Clinically there is no effective therapy to treat them. Physicians are treating colon to liver metastasis either like a colon cancer or as a liver cancer, and there is no clear consensus.

Causal network (CN) is a stricter form of BN. Constrained by Markov conditions among variables, CN reduces the ambiguous directions of edges thus more clearly reveal the real causations among variables, which could be critical during gene regulatory network (GRN) reconstruction. Here we proposed a searching algorithm 'FFBN', which is derived from the famous FGS algorithm [59] used for searching the optimal model from a Causal network. Compared to FGS, 'FFBN' optimized the converging criteria and reduced unnecessary steps during model searching thus speeded up the whole reconstruction process.

In this study, FFBN is applied on reconstructing GRNs for three sets of transcriptome samples: primary colon cancer (PC), primary liver cancer (PL), and

colon to liver metastasis (CLM) to build GRNs. By comparing three regulatory networks, top hubs that are uniquely existed in CLM network, commonly existed hubs between CLM-PC and CLM-PL respectively, are identified. After mapping hubs and their neighbors onto DAVID database [60], we observed a series of enriched pathways that may highly involve in molecular mechanisms related to colon to liver metabolisms. On top of that, several hub genes that highly regulates various members in the pathways as potential drug targets are selected.

3.2 Materials and Methods

3.2.1 Data availability

In the study, all the microarray data (gene expression) is obtained from gene omnibus (GEO) database [61] of national center for biotechnology institute (NCBI). Since metastatic cancer study is relatively rare on GEO database, to make use of all possible datasets, a semi-automatic searching approach has been conducted: description pages (HTML) of all 67,362 datasets (May, 10th, 2016) from GEO are downloaded and traversed for searching title, summary and description paragraphs with key words “metastasis”, “metastatic”, “metastases” and platform constraints “AFFY”. Then from the remaining datasets, qualified datasets are selected with the following constraints: liver metastasis, colon cancer, liver cancer; tissue sample; AFFY U133A (GPL96) or AFFY U133PLUS2.0 (GPL570). Eventually, a subset of 9 datasets are considered qualified (Table 1). All the samples within are further classified into primary liver set (378 samples), primary colon set (186 samples) and colon to liver metastasis set (107 samples).

3.2.2 FFBN algorithm

FGS algorithm FGS is a score-based algorithm which takes advantage of D-separation concept and generates an optimal network structure. FGS consists of two phases: a ‘forward phase’ and a ‘backward phase’. Starting from an empty network, the ‘forward phase’ iteratively tries to add exactly one edge that has the highest ‘Bayesian information criterion’ (BIC) score if this addition doesn’t create a cycle. The orientation of this edge is randomly given. This phase ends if no more edge addition can lead to an increase of BIC score. ‘Backward phase’ takes over the result from ‘forward phase’ and tries to remove one edge at a time if the removal will not cause a score decrease. This phase terminates when no more edges can be removed from the network. Importantly, either an edge has been added or removed, the whole network would be converted back to a pattern, which means any edges that are not connected to the sink node in D-separated structures would be unoriented. Then Meek’s rules would be applied to re-orient some edges. In this process, unoriented edges might become oriented and incorrectly oriented edges might be corrected.

Meek’s rule In the FGS algorithm, a series of rules called Meek’s rules (Figure 4) [62] are applied to orient some of the undirected edges to convert the network to so-called ‘pattern’ network.

FFBN algorithm In FGS algorithm, the computation burden mainly lies in two parts: calculating BIC score for the new network structure and re-orienting edges with Meek’s rule at each step. Calculating BIC score is inevitable, so we developed FFBN algorithm that optimizes the re-orientation process (Displayed in Table 2). Instead of

applying R1-R4 after each edge insertion or deletion, FFBN applies Meek's rules twice in the entire process, one after 'forward phase' and one after 'backward phase'. This optimization significantly reduces the time complexity of the search process.

As mentioned in section 2.2.1, D-separation can be used to quickly identify the orientations after determining the correlations between variables. In FFBN, conditional correlation is used to determine the correlation and then D-separation concept is applied to orient edges. For example, consider an edge $X \to Y$ to be inserted into the network, where Y has a parent set [29]. FFBN then calculates the following probability $\mathrm{P}(Y|Z,X) - \mathrm{P}(Y|Z)$, $\mathrm{P}(Y|W,X) - P(Y|W)$ and $\mathrm{P}(Y|X,Z,W) - P(Y|Z,W)$. If $\mathrm{P}(Y|Z,X) - \mathrm{P}(Y|Z)$ is the highest positive value, then clearly X and Z together better explains Y in the model than separated. Thus Y should be the collider of both X and Z. D-separation is then used to orient $X - Y - Z$ to $X \to Y \leftarrow Z$.

There are several parameters that control the density of the network: P is the penalty term in the BIC score formula. The lower the P is, the more likely edges are to exist in the network; D is the maximum degree of any node. This constraint is to prevent some specific nodes from becoming super hub nodes due to false collateral correlations between these nodes and the rest of the nodes. To add or delete an edge from the G, the BIC score must increase.

3.3 Results

3.3.1 Benchmark results

DREAM challenge data [55] provides a series of gene expression datasets with golden standards, which are curated through biological experiment, thus highly reliable.

Hence, DREAM5 data is used as input to benchmark the performance of FFBN and FGS. The winner of the DREAM5 challenge: GENIE3 [63] is also incorporated into the benchmark as performance reference. DREAM5 consists of four networks, *in silico* network, *S.aureus* network, *E.coli* network and *S.cerevisiae* network. Each of which contains 1,643, 2,810, 4,511, 5,950 variables, respectively. The *in silico* and *S. cerevisiae* network, which is the smallest and largest network in DREAM5 dataset, are used in the benchmark for evaluating the performance of the three methods.

Table 3 listed the benchmark result of three algorithms with two networks in DREAM5 dataset: For the smaller *in silico* network, GENIE3 provides the best AUROC (0.82) and AUPR (0.3) while has the highest speed among the three (Table 3(a)). FFBN maintains at least equally good AUROC and AUPR compared to FGS under different all parameters while offers an 11% to 75% faster searching speed. For S. cerevisiae network, three methods have an equal AUROC (~0.52) while FFBN and FGS have better AUPR (0.07) than GENIE3 (0.02) (Table 3(a)). Speed wise, GENIE3 still maintains the fastest speed. FFBN provides 19% to 96% speed increase compared to FGS (Table 3(b)).

3.3.2 Constructing GRNs for liver cancer, colon cancer and colon to liver metastasis using FFBN on a whole transcriptome scale

In total 14,213 common genes are included for reconstructing the GRN for three cancer subtypes. FFBN and FGS algorithm are executed on the supercomputer ‘Carbonate’ at Indiana University with 16 cores and 100 GB RAM. Table 4 shows the total summary of constructed regulatory networks for CLM, PL and PC groups

respectively. After the GRN construction, 12,087 overlapped genes (nodes) are included among PL, PC, and CLM. In this study, any genes that have out-degree larger than 10 are defined as hub genes.

3.3.3. Hub genes matching with oncoKB and Functional comparisons among subnetworks of CLM, PL and PC via pathways

Hub genes regulate many other genes, thus play important roles in biological networks. A change of hub gene (mutations, expression level change, etc.) could significantly influence pathways and biological functions. In total, there are 118, 522 and 223 hub genes existed in CLM, PL and PC network respectively. These hub genes across three cancer types are firstly mapped onto onco-KB [64], which is a knowledge base for all oncogenes supported by clinical evidence. Among all hub genes, 11, 44 and 19 genes occur in onco-KB for CLM, PL and PC network respectively. Representative subnetworks for CLM, PL and PC respectively are then selected based on their unique CLM hubs, CLM-PL common hubs and CLM-PC common hubs indicated in Figure 5. Numbers in this figure indicates the number of hub genes (out-degree > 10). There are 70 hubs unique to CLM, 13 common hubs between CLM and PL, and 25 common hubs between CLM and PC. They represent the unique subnetworks of CLM, the common subnetworks between CLM and PL, and common subnetworks between CLM and PC, respectively. All genes within subnetworks are then mapped onto DAVID for identifying enriched pathways.

3.3.4 Enriched pathways of unique CLM subnetworks

Ten pathways were statistically significantly enriched among the 70 unique CLM subnetworks, P-value<0.05. Besides five pathways that are not obviously cancer related (e.g. Alzheimer's disease pathway), the other five pathways represent unique aspects of the CLM malignant process. Among them, Metabolic pathway contains the largest number of enriched genes in the subnetworks. PHB is the largest hub gene with five neighbors: NME2, GSTZ1, CYP4F2, G6PC3 and ATP5G1 as representative actionable drug targets. Overexpression of PHB results in the up regulation of NME2 and G6PC3, which plays a critical role in the metabolic pathway.

Figure 6 lists four representative enriched pathways by CLM-unique hub genes and their neighbor genes. In this figure, Red dots are hub genes in CLM unique subnetwork while others are non-hub genes in the subnetwork; Green lines indicate the regulating directions. Proteasome pathway (Figure 6) is essential for various cellular processes, such as cell cycle regulation, which may cause oncogene mutations. It is also an important part of Ubiquitin-Proteasome Systems (UPS), which promotes cancer metastasis by regulating epithelial to mesenchymal transition (EMT) [65]. In insulin pathway (Figure 6), EMT modulates the secretion of insulin, and hyperinsulinemia is a risk factor of colon cancer and liver cancer [66]. SH2B2, a hub gene in the CLM network, regulates more genes in the insulin pathway than any other gene. SH2B2 plays a role in signal transduction from the receptor to the Shc/Grb2 pathway and is highly related to the insulin pathway.

GnRH pathway (Figure 6) is responsible for the release of follicle stimulating hormone and luteinizing hormone from the anterior pituitary. Activation of the GnRH signaling pathway can activate several motigen-activated protein kinases (MAPKs), including extracellular-signal-regulated kinase (ERK), Jun N-terminal kinase (JNK) and p38 MAPK [65]. Notably, gene LHB regulates more genes in GnRH pathway than any other genes. It encodes the beta subunit of luteinizing hormone, which is controlled by GnRH.

AMP-activated Protein Kinase (AMPK) pathway coordinates cell growth, autophagy and metabolism [67]. Recent studies indicate that the activation of the whole pathway can suppress tumor development and progression by regulating inflammation and metabolism [68]. In enriched AMPK pathway (Figure 6), 12 hub genes outside the pathway regulate 12 individual downstream genes involved in the pathway.

3.3.5 Enriched pathways of CLM-PC common subnetworks

Figure 7 lists two pathways that are commonly enriched by CLM-PC subnetworks. In this Figure, Blue lines, red lines and gray dash lines indicate regulations in PC, CLM and both networks respectively. Highlighted dots indicate hubs in the regulatory networks. Notably, in T-cell receptor signaling pathway (Figure 7), a large proportion of regulations (gray dashed line) are overlapped and the majority of them are from one gene, VAV3, which is a common hub in both CLM and PC subnetworks and regulates 28 common downstream genes, showing the expression consistency. This gene, along with its upstream genes, CFTR, SESN1 and LY6G6F,

are overexpressed in both CLM and PC compared to PL. Moreover, TRBC1 has the most regulations in the pathway, indicating its great regulatory impact on the pathway in both cases. Taken together, CLM and PC share a certain degree of similarity in terms of T-cell based immune response, given the numerous regulators and downstream genes that are present in both CLM and PC subnetwork.

3.3.6 Enriched pathways of CLM-PL common subnetworks

After ruling out non-cancer related pathways, remaining pathways are then classified them into two categories: immune response related pathways, including ‘phagosome’ and ‘antigen processing and presentation’ and metabolic related pathways, including ’endocytosis’, ‘oxidative phosphorylation’ and ‘ribosome’. Among them, Oxidative phosphorylation (OXPHOS) pathway plays different roles in various cancer types. Normally, cancer cells have upregulated aerobic glycolysis compared to normal cells, with a concomitant downregulation of OXPHOS activity. Cancer cells prefer glycolysis over oxidative phosphorylation because the additional metabolites of glycolysis may benefit cell proliferation [69]. This phenomenon is called ‘The Warburg Effect’ and has been observed in many cancer types, including liver cancer. However, in many recent studies, a reversed ‘Warburg Effect’ -low regulation of glycolysis instead of high regulation of OXPHOS - have been found in certain cancer cells, including leukemia, lymphoma, pancreatic ductal adenocarcinoma and primary colon cancer [70, 71]. Figure 8 indicates the different regulations between CLM and PL subnetwork in the commonly enriched OXPHOS pathway. Blue lines, red lines and gray dash lines indicate regulations in PL, CLM and both networks, respectively. Highlighted dots

indicate the common hub genes existing in PL and CLM. Interestingly, there is no common regulation (grey line) between these two subnetworks, which implies that CLM and PL have unique regulation mechanisms towards OXPHOS pathways.

3.4 Discussion

Human genome contains ~25,000 genes while Onco-KB covers 1,045 genes. Hence chi-square tests are performed to see if the frequency of hub genes that fall into onco-KB is significantly higher than background. All three networks show significantly higher frequencies with $p<0.005$, $p<0.001$ and $p<0.012$ for CLM, PL and PC respectively. This indicate that compared to randomly selected genes, hub genes are much more linked to cancer in three networks. This indicates the value of further investigating hub genes and their neighbors.

The analysis results suggest that CLM has its own self-promoting and self-regulating mechanisms via various pathways and core genes. For instance, GnRH pathway is enriched in the CLM unique subnetwork. Previous studies have shown that if GnRH receptor is expressed in cancer cells, then administering GnRH analogs reduces cancer proliferation and metastasis *in vivo* [72]. The results indicate that further investigation of the possibility of treating CLM with GnRH analogs may be worthwhile. In another enriched pathway ‘metabolic pathway’, Notably, a hub gene PHB regulates many targets, including NME2 and G6PC3. Both PHB and its targets NME2 and G6PC3 are over-expressed in CLM compared to PC. PHB acts as a trans-acting regulatory RNA and plays a role in human cellular senescence and tumor suppression.

CLM and PL share certain GRN subnetworks related to immune response and metabolic pathways. Notably, the similarity is also present when considering the oxidative phosphorylation pathway, where hub gene ATP5B and its neighbor genes show consistent under-expression in CLM and PL compared to PC. This phenomenon clearly indicates the similarity between CLM and PL on energy generation. However, there are no common regulations between CLM and PL, indicating the existence of different inhibitory mechanisms towards this pathway.

CLM and PC share similar immune responses. In T-cell receptor pathways, one common hub gene between CLM and PC is VAV3. VAV3 acts as an exchange factor for GTPases. This implies that CLM and PC share a very similar T-cell receptor-related immune response via VAV3.

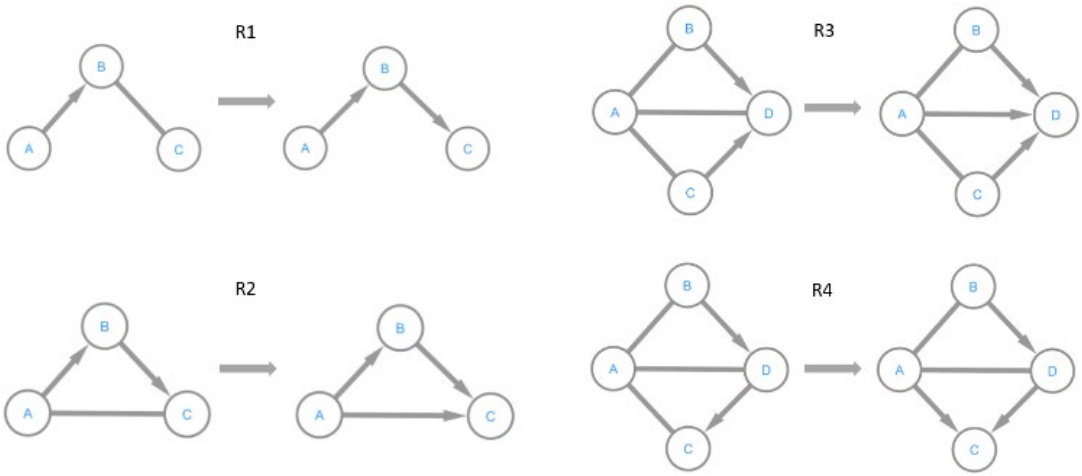

Figure 4 Meek's rules

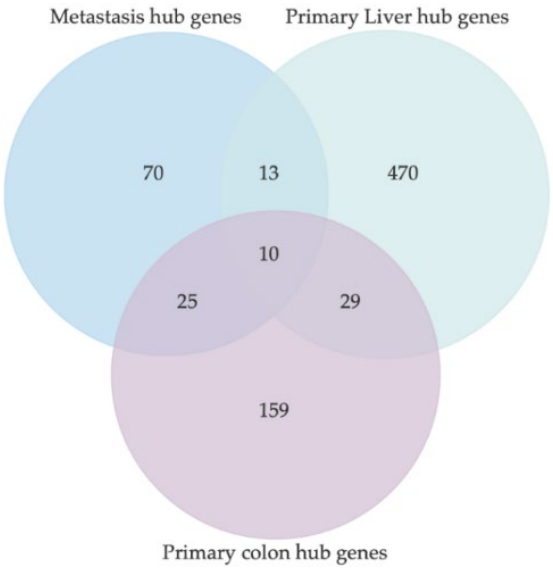

Figure 5 Overlap of hub genes for the CLM, PL, and PC regulatory networks

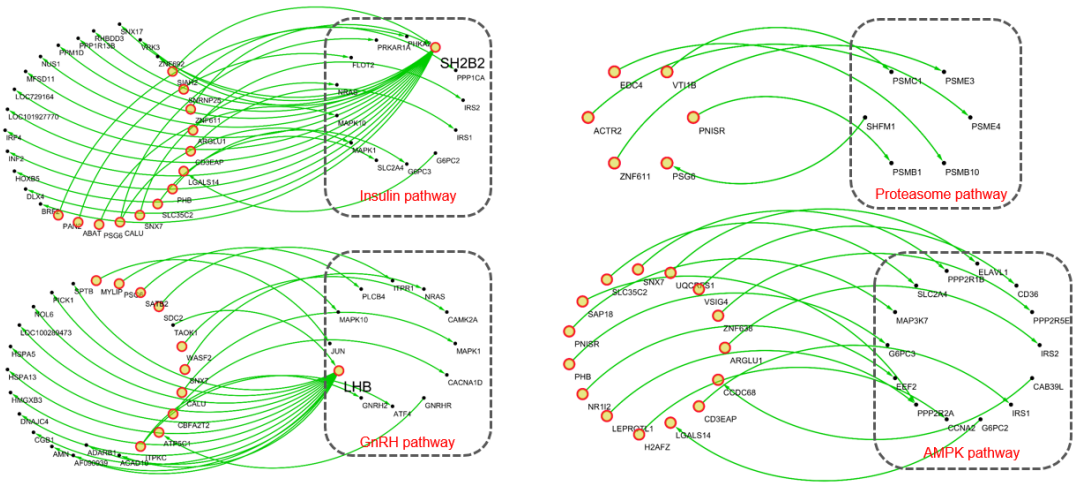

Figure 6 Four representative pathways in CLM unique subnetwork

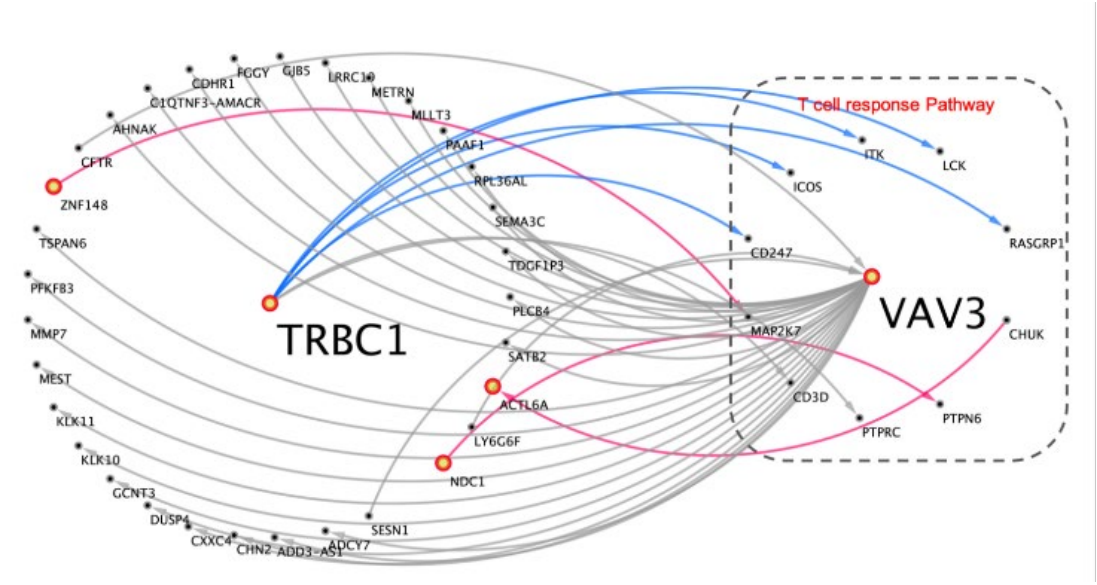

Figure 7 Regulations among genes involved in the T cell response pathway

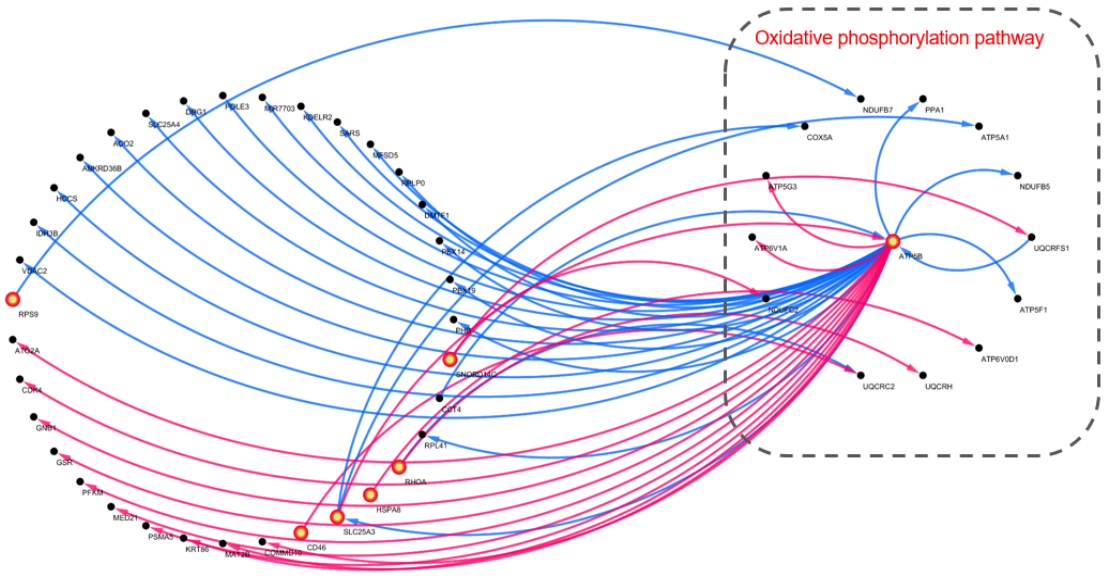

Figure 8 Regulations among genes in oxidative phosphorylation pathway

Table 1 Data composition and sources

Subtype	CLM	PL	PC
Session ID	GSE41258 GSE49355 GSE68468	GSE45436 GSE51401 GSE55092 GSE6764 GSE17548 GSE19665 GSE60502 GSE14243 GSE9843	GSE41258
Sample Size	107	378	186

Table 2 FFBN algorithm description

Algorithm **FFBN** (*M*):
Input: a matrix *M* of *N* variables and *S* samples
Output: a directed acyclic graph (DAG) *G*
Parameters: BIC penalty *P*, maximum depth *D*
1. Calculate pairwise covariance matrix M_c for *N* variables;
2. Calculate Bayesian information criterion BIC for all possible edges from *G*;
3. Build set *E*=$\{e_1, \ldots e_n\}$, where each *e* is an edge causing positive **BIC** difference;
4. **WHILE** E is not empty:
5. **IF** adding *e* result in no cycles and related nodes' degree<*D*:
6. *G*=*G*+*e*;
7. *E*=*E*-*e*;
8. Re-calculate **BIC** (*G*-*e, P*)-**BIC**(*G, P*) for each $e \in E$ from updated *G*;
9. *G*=**Adjust_Direction**(*nt*, G);
10. *G*=**Convert_to_Pattern**(*G*);
11. **FOR** each $e \in G$:
12. **IF BIC**(*G*-*e, P*)-**BIC**(*G, P*)>0:
13. *G*=*G*-*e*;
14. Re-calculate **BIC** (*G*-*e, P*)-**BIC**(*G, P*) for each $e \in G$;
15. *G*=**Adjust_Direction**(*nt*, G);
16. *G*=**Convert_to_Pattern**(*G*);
17. **RETURN** *G*;

Function **Adjust_Direction** (*G*):
Input: a graph *G* and target node *nt* in an inserted edge
Output: a graph with adjusted direction G'
1. **FOR** each edges e pointing to *nt*:
2. **IF** *e* is undirected:
3. Orient *e* pointing to *nt*;
4. **RETURN** *G*;

Function **Convert_to_Pattern** (*G*):

Input: a graph *G*

Output: a pattern graph *G'*

1. Unoriented edges not involved in any v-structure in *G*;
2. Apply Meek's rule R1, R2, R3 and R4.
3. **RETURN** *G*;

Function **BIC** (*G, P*):

Input: a graph *G* and penalty *P*

Output: BIC score

1. $\text{BIC score} = P\ln(n)\,k - 2\ln(\hat{L})$;

2. **RETURN** BIC score;

Where:

P is the penalty term of BIC;

n is the number of samples for each variable';

nt is the target node of a directed edge.

k is the number of nodes in the current network G;

$\hat{L}$ is the likelihood of a specific network structure.

$\hat{L} = \prod_{i=1}^{n} p\left(x_i, x_j \middle| pa_j^G, D\right)$ Where each $p\left(x_i, x_j \middle| pa_j^G, D\right)$ is the probability between node x_i and x_j given its parents set pa_j^G and dataset D

Table 3 Benchmark results of FFBN and FGS

3 (a) accuracy comparisons

Accuracy	AUROC *In silico*	AUPR *In silico*	AUROC Yeast	AUPR Yeast
FFBN	0.74	0.26	0.51	0.07
FGS	0.72	0.25	0.52	0.07
GENIE3	0.82	0.3	0.52	0.02

3 (b) runtime comparisons. P indicates the penalty parameter used in calculating BIC. T indicate the trees generated for each gene in GENIE3

Runtime (minutes)	P=10	P=6	P=4
In silico network			
FFBN	1.8	54.3	232.3
FGS	2.0	67.2	404.5
GENIE3	0.6 (T=1)	15.9 (T=100)	121.7 (T=1000)
Yeast network			
FFBN	121.2	1223.3	2022.5
FGS	144.6	1921.8	3914.2
GENIE3	1.1 (T=1)	23.9 (T=100)	619.0 (T=1000)

Table 4 Summary of three generated GRNs

Type	Nodes	Edges	FFBN Runtime (mins)	FGS Runtime (mins)
CLM	13,429	31,140	2,430	4,537
PL	11,139	27,092	5,931	NA
PC	11,273	17,091	4,299	8,399

4. SCN: Spectral Clustering for Network Based Ranking to Reveal Potential Drug Targets and Its Application in Pancreatic Ductal Adenocarcinoma

4.1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic malignancy. Due to its wide heterogeneity, PDAC acts aggressively and responds poorly to most chemotherapies, causing an urgent need for developing new therapeutic strategies. Cell lines have been used as the foundation for drug development and disease modeling. CRISPR-Cas9 plays as a key tool for every step-in drug discovery: from target identification and validation to preclinical cancer cell testing. Using cell-line models and CRISPR-Cas9 technology together makes drug targets prediction feasible. However, there is still a big gap between prediction results and actionable targets in real tumor.

Network based analysis have greatly benefited cancer biology. So far, biological network has been used widely in numerous studies for identifying genes related to a certain therapy through curated database, specialized drug-protein [73] or protein-disease networks [74, 75]. (1) Curated databases, such as STRING protein-protein interaction [19] network and KEGG [76] for pathway network, can provide complete genome wide networks that contains entire gene regulations, signal transductions and gene protein associations. However, these types of methods are not built for specific cancer types, making them too generalized. And it is also hard for people to analyze them as a whole. (2) drug-protein network is often used to investigate the mechanism of drug action and drug target prioritization [77]. For

instance, Isik et.al provided drug target identification by perturbed gene expression from Connectivity Map (CMAP) [56] and protein-protein Interaction (PPI) network information. However, these technologies did not directly connect drug with disease genes. (3) constructing protein-disease networks is another approach to identify genes-disease associations for selecting therapeutic targets in cancer [78]. *Ferrero* et al. proposed a semi-supervised network approach, which evaluates disease association evidence and make de novo predictions of potential therapeutic targets based on that [79]. These types of method fail to incorporate target information into their models to accurately predict drug targets.

To better prioritize drug targets for PDAC, a method called 'SCN' that systematically utilizes expression data from tissue and cell-line, along with gene perturbation data and PPI network is proposed to select and rank druggable targets that effectively works on tissues. Systematically network structure comparing among the tumors and cell line models is used to discover the gene functional transfer among different models, and it is important to choose the most appropriate targets for patients. By connecting knockout of genome in cancer cell and cell death phenotype to observe patients network co-variation and select the optimum genes as potential targets. To our knowledge, this is the first-time people proposed a model to systematically predict potential targets by considering subnetwork similarity between tumors network and cell line network. On the other hand, ranked drug targets are validated by 1) mapping them onto existing PDAC drug targets; 2) applying pathway analysis on drug targets and the

clusters there are within, to show their functional associations with PDAC; 3) performing survival analysis for top ranked drug targets.

4.2 Materials and Methods

4.2.1 Materials

Summary of all omics data involved in this studied can found in Table 5.

Expression data of PDAC In total, expression data that is gathered from 263 samples across three groups are used in this study, including 92 PDAC cell-line samples, 113 PDAC tissue samples and 58 adjacent normal pancreas tissue samples. These data are all from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database and all generated from Affymetrix Human Genome U133 Plus 2.0 Array, which contains 54,675 probes pointing to over 20,000 genes.

Protein-Protein Interaction network STRING [19] is a comprehensive and public pathway database (https://string-db.org/), which accumulates numerous prior knowledge of biological pathways and protein-protein interactions. STRING network protein links version 11 data is also included in this study.

Genome wide CRISPR-Cas9 screening data and gene essentiality value To measure gene essentiality, CRISPR-Cas9 v3.3.8 screening data from 'Project Achilles' [80-82], (https://portals.broadinstitute.org/achilles) which includes genome wide CRISPR-Cas9 screening data that affect cell survival across 43 tumorous cell lines and genome wide RNAi screening data over 501 cell-lines is utilized in this study. We choose CRISPR-Cas9 over RNAi is because recent study has indicated that

compared to RNAi, CRISPR-Cas9 method has been proven to have less off-target effects, thus better for cancer drug-target related research [83]. In total, gene perturbation data of 74,222 sgRNAs on 17,733 genes across 22 PDAC cell-lines are included in this study.

FDA approved drug targets All FDA approved drug and their targets are downloaded from Drug bank [5]. In total, all targets have been mapped onto 1,317 genes, of which 283 genes are cancer drug targets.

4.2.3 Methods

SCN takes multiple types of omics data from tissue and cell-line data as input to rank druggable targets to recommend potential targets. SCN mainly consists of four steps (Figure 9): (**A**) Constructing integrated tissue PDAC network; (**B**) Constructing weight integrated gene perturbation network of PDAC cell by cells gene expression profiles and CRISPR-Cas9 genome perturbation; (**C**) Spectral clustering for subnetwork partition; (**D**) Graph structure similarity alignment between disease genes in tumors and gene perturbation networks in cancer cells and score to priority potential targets.

STEP A: Construct an integrated network for tissue PDAC. The algorithm first compares tumor tissue and normal tissue expression profile to select overexpressed genes existed in tumor. Since the sample number of tissue tumor and normal groups are not equal, an unpaired T-test is performed while the corresponding p-value cut-off 0.05 is used. Log fold changes between tumor and normal tissue samples are calculated for all significantly overexpressed genes. The algorithm then constructed a

correlation network by calculating Pearson correlation coefficient as edge weights. Log fold change is then used as the node weights in the network. The algorithm then maps integrated network onto STRING PPI network and selected the overlapped subnetwork. The rationale of mapping is that: 1. we believe high correlations among genes that also reflect on protein level are more likely to be true; 2. mapping both tissue integrated network and cell-line tissue integrated network onto the same PPI network makes them comparable via the PPI network. Eventually, a network with skeleton from PPI network, edge weights from pair-wise gene correlation, and node weights from Tumor -versus-Normal log fold change is constructed.

STEP B: Construct an integrated perturbation network of pancreatic cancer cells. Only genes that are selected in STEP 1 are picked from cell-line expression profile for integrated network construction. Similarly, pair-wise Pearson correlation coefficient of these genes are calculated to build a correlation network. The network is then mapped onto STRING PPI network and only the overlapped subnetwork is kept. Then gene essentiality value (gene knockout data) generated by CRISPR-Cas9 is then integrated into the network as node (gene) weights. Finally, two constructed networks share the same nodes and edges but with totally different node weights and edge weights.

STEP C: Dimension reduction and network partition. In the network partition process, high dimension network is reduced to low dimension clusters thus common features among variables can be better captured from a graph perspective. Given a

graph G with n nodes and k categories, the objective function of spectral clustering can be described as:

$$min.\,cut(A_1, \dots A_k) = \frac{1}{2}\sum_{i=1}^{k} W(A_i, \overline{A_i}) \quad (16)$$

Where $W(A_i, \overline{A_i})$ is the weight between cluster A_i and its complement set $\overline{A_i}$. This is well-known as RatioCut [84] problem and has been proven as an NP-hard discrete problem. To approximate the optimal solution, Spectral clustering [85] is applied to divide a network into clusters based on the spectrum (eigenvalues) of the corresponding similarity matrix.

The general steps of performing spectral clustering can be described as:

1. For n variables (nodes), construct an affinity matrix S

$$S=\begin{pmatrix} 0 & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & 0 \end{pmatrix} \quad (17)$$

Where Sab in the matrix indicate the connectivity between variable a and b in the network.

2. Construct a diagonal matrix D as degree matrix

$$D = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{pmatrix} \quad (18)$$

Where d_a in the matrix indicate the degree (total edges) of variable a in the network.

$$\text{Clearly, } d_a = \sum_{k=1}^{n} S_{ak} \quad (19)$$

3. Construct Laplacian matrix $L' = D - S$ (20)

4. Normalize the Laplacian matrix $L = D^{-\frac{1}{2}} L' D^{-\frac{1}{2}}$ (21)

5. Perform singular value decomposition for matrix L

6. Pick top K eigenvalues and their corresponding eigenvectors to generate a $N * K$ matrix
7. Perform K-means clustering [86] on the extracted matrix.

Clearly, the Laplacian matrix L consists of two types of node information: local information, which is node connectivity towards its neighbors in matrix S; global information, which is node degrees, or 'influence' towards the entire network. Hence, the clustering can be considered as selecting similar nodes based on their local and global similarities. Inspired by this idea, the Pearson Correlation Coefficient (CC) among nodes instead of the connectivity value (0 or 1) is used to fill up affinity matrix to measure the local similarities among genes. We also plugged log fold change of tumor versus normal expression value in degree matrix to indicate the global influence of genes. Hence, matrix S and D becomes:

$$S' = \begin{pmatrix} 0 & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & 0 \end{pmatrix} (22) \text{ and } D' = \begin{pmatrix} FC_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & FC_n \end{pmatrix} (23)$$

Where r_{ab} is the CC between gene a and b in tumor expression profile, and FC_a is the log fold change of gene when comparing its expression value in tumor group to value in normal group. And to fulfill formula (19), S needs to be normalized to:

$$S'' = \begin{pmatrix} 0 & \cdots & \frac{abs(r_{1n})FC_1}{\sum_{k=1}^{n} abs(r_{1k})} \\ \vdots & \ddots & \vdots \\ \frac{r_{n1}abs(FC_n)}{\sum_{k=1}^{n} abs(r_{nk})} & \cdots & 0 \end{pmatrix} (24)$$

Hence, the final Laplacian becomes $L' = D' - S'' (25)$.

For K-means clustering, picking the optimal K could be arbitrary. In this study, K is equal to the number of eigenvalues that the algorithm picked. Too many or too few

eigenvalues will result in overfitting and underfitting respectively. Hence, an intuitive approach is applied: from K=1 to the number of total number of variables, K-means algorithm is performed while Hartigan's number, which is a measurement of the clustering quality by comparing two clustering results, is calculated. For a K-means clustering, if the number is greater than 10, than having K+1-means clustering is of value [87]. The first K that result in the Hartigan's number less than 10 will be selected, even though this scheme of picking K does not guarantee global optima.

STEP D: Graph structure similarity alignment between subnetworks of dysregulation genes in tumors and perturbation networks in cancer cells and score to rank for priority potential targets. Spectral clustering is applied on tissue integrated network to look for genes that show common features. 1,317 targets (genes) for all FDA approved drugs are then mapped onto clusters. Then for the successfully mapped drug targets, the influence that each target might cause towards their clusters are evaluated. The assumption is that a drug target's 'influence' is limited in its cluster. In that case, a drug target's influence towards any node is determined by the paths between them. Hence, given a graph G (V, E), where V and E are node set and edge set. Assume the node weight set is W and edge weight set is Y. For a drug target x, its maximum 'influence' towards all other nodes can be described as:

$$\sum_{k \in V} W_k \prod_{i=k}^{x} Y_{i,i_{next}} \quad (26)$$

$$where\ Y_{i,i\_next} \in \mathrm{E}$$

Where $\prod_{i=k}^{x} Y_{i,i\_next}$ indicates the transmitted influence from target x to a node k via one possible path. Obviously, to maximize term (26), for every other node i, we need to find the most correlated path between x and i. Thus, the total influence of x becomes:

$$\boldsymbol{TI} = \sum_{k \in V} W_k \max\left(\prod_{i=k}^{x} Y_{i,i_{next}}\right) \quad (27)$$

Here, the term $\max\left(\prod_{i=k}^{x} Y_{i,i_{next}}\right)$ represents the most correlated path between x and i. And we define term (27) as Target Influence score (TI). Subsequently, a scoring scheme is developed for calculating TI scores for all 367 drug targets.

4.3 Results

4.3.1 Potential target subnetworks and targets for PDAC by SCN algorithm identification

Overlapping 15,664 common genes among 263 gene expression profiles tumor tissue, normal tissue and cell-line are included for SCN analysis, among which are 7,376 genes are significantly dysregulated by non-paired t-test where p-value is less than 0.05. 4584 genes out of 7,376 genes are significantly over-expressed in tumor tissues group compared to normal tissue group. All 4,584 genes are then mapped onto STRING human PPI network. 4,144 genes have overlapped with PPI network. 367 out of 4,144 are drug targets of FDA approved drugs. In total, 4,141 genes and associated 931,288 pairs of gene-gene interaction of network with 367 FDA approved drugs' targets (which includes 90 cancer drug targets) are input into SCN algorithm to seek potential targets for PDAC patients.

In total, 198 subnetworks (clusters) have been identified by SCN algorithm for PDAC patients. All 367 targets are scored and ranked by SCN system. Table 6 shows

the top ten targets and two well-known PDAC drug targets ranked by ‘SCN’. Column 2 indicates the Rank predicted from SCN; Column 3 indicates cancer drug target information; Column 4 lists the average expression values in tumor tissue samples; Columns 5 lists average expression values in normal tissue samples; Column 6 indicates log2 fold change of expression differences between tumor group and tissue group; Column 7 indicates T value from t test between tumor and normal group; Column 8 indicates P-value from T test between tumor and normal group. Column 9 indicates gene essentiality value (Cell survival rate at T3 versus at T0). Positive values and Negative values indicate an enhanced and reduced cell survival rate respectively in vitro. of which POLE2 and DHFR are known cancer drug targets. ERBB2 and MTOR are PDAC drug targets.

The 12 selected genes are all highly expressed in tumor tissue compared to normal tissue. Moreover, the loss of all 12 genes cause a reduced cell survival. Among them, two widely accepted targets ERBB2 and MTOR in treating PDAC are caught by SCN algorithm. PGK1, POLE2 and HMMR are the top three ranked targets. PGK1 is in a cluster of 41 genes. POLE2 and HMMR are together in a cluster 67 genes. Figure 10 shows the expression level of two clusters containing top three ranked targets in tumor tissue, normal tissue and cell-lines. It can be observed that these genes show a concordant high expression pattern in cell-line and tumor than normal group.

Figure 11 displays the compositions of the subnetworks of PGK1, POLE2 and HMMR as well as their expression pattern in tumor, normal and cell-line models. Cube node indicates known targets while circle node indicates other genes. Red and blue

lines indicate positive and negative correlations, respectively. Line shade indicates correlation intensity. Nodes are placed in a clockwise order by their ascending degrees. Subgraph (A) indicates the top ranked Drug targets ‘PGK1’ and the subnetwork of its cluster. Subgraph (B) indicates the second and third ranked Drug targets ‘POLE2’, ‘HMMR’ and the corresponding subnetwork of their common cluster. Yellow highlighted genes are common genes between HMMR and POLE2. Subgraph (C), (D) and (E) are overall survival results from cox regression to test the top three ranked drug targets PGK1, HMMR and POLE2 respectively.

Glycolytic enzyme phosphoglycerate kinase 1 (PGK1) is a gene that is coding for a glycolytic enzyme that catalyzes synthesis of 3-phosphoglycerate. Its functions and mechanisms haven’t been completely understood. As an inhibitor, PGK1 inhibits the secretion of vascular endothelial growth factor (VEGF) and interleukin-8, thus inhibits Angiogenesis [88]. However, multiple studies have suggested that in metastatic tumor cells, PGK1 plays a completely contrary role. Overexpression of PGK1 facilitates not only tumor growth and interaction with microenvironment, but tumor invasion and metastasis in liver, gastric and prostate cancer [89, 90]. In this study, PGK1 has been identified as the target that can cause the highest influence towards its cluster (shown in Figure 11(A)). It interacts not only with the greatest number of genes, but also with the greatest number of other targets in the cluster. Most of its correlations with its neighbors are positive.

DNA Polymerase Epsilon 2, Accessory Subunit (POLE2) is high involved in DNA repair and replication. It has been previous reported to have high association with

colorectal cancer [91]. In this study, POLE2 is ranked as the second highest target. Even though its cluster is much larger than the cluster of PGK1 (shown in Figure 11(B)), the influence of POLE2 towards the whole cluster is not as strong as influence of PGK1.

Hyaluronan Mediated Motility Receptor (HMMR), which is target with third highest score, is highly involved in cell motility. HMMR forms a complex with BRCA1 and BRCA2, thus it has been identified as a high-risk factor in multiple cancer types such as breast cancer and fibrosarcoma [92, 93]. Interestingly, HMMR is in the same cluster with POLE2 (shown in Figure 4-5). Their degrees and ranks are very similar, implying their equally influence towards the whole cluster.

4.3.2 Pathway enrichment analysis for top three ranked targets and their clusters

For all 198 clusters, 'Gene Set Enrichment Analysis' (GSEA) are performed for getting enriched pathways [94]. 'C5 go gene sets BP GO biological process' database version 6.2, which contains 4,436 gene sets annotated by GO term with their functions, are selected as reference pathway database while significance level $P<0.05$ are used in pathway analysis. GSEA analysis required a ranked gene list to perform such analysis, so we used log fold change of tumor vs normal tissue as their weights and ranked them.

The top ranked gene, PGK1 with its cluster, has significantly enriched 'CARBOHYDRATE_CATABOLIC_PROCESS'. The second and third gene, HMMR and POLE2, with their cluster, have significantly enriched multiple pathways such as 'CELL CYCLE' and 'MIOTOSIS'. These pathways are all highly related to cell cycle and cell division, suggesting these two genes along with their cluster members, are

critical components in regulation cell cycles. Moreover, HMMR and POLE2 enriched 8 pathways of 11 total enriched pathways that are enriched by the entire cluster, suggesting common functional activities.

4.3.3 Target selection validation by clinical outcomes

Survival analysis for differentially expressed genes PGK1, HMMR and POLE2 from public database 'GEPIA' (http://gepia.cancer-pku.cn/) are performed. GEPIA [95] is a public database containing 9,736 tumors and 8,857 normal samples from TCGA [96] and GTEx [97] projects. All three targets showed significant difference (Hazard ratio P-value<0.01) in patients' survival (Figure 11 (c), (d), (e)) Low expression of these three genes provide significantly higher survival than high expression. Survival curves of all three genes show a similar pattern at around 20th months, in which low expression curves start to have a clear segregation from high expression curves.

4.3.4 Targets accordance comparison between clinical drug treatment in pancreatic cancer and selection by SCN algorithm

Amanam and Chung systematically investigated all currently available targeted therapies and drug targets for pancreatic cancer [98]. Los of studies have reported HER2 overexpression in up to 45% of patients with PDAC [99]. This is because HER2 amplifications often occurs in PDAC [100]. Known drug targets are mapped to the ranks system and listed result in Table 6.

In this study, HER2 are ranked 14th by SCN. SCN algorithm covered four out of top 100 only except to MAP2K1, of which ERBB2 and MTOR are ranked relatively

high (rank 14 and rank 32 respectively). All the missing targets are not included in 4,414 genes for constructing integrated networks in the beginning.

4.4 Discussion

Tumor cells prefer glycolysis to oxidative phosphorylation for providing energy during proliferation and metastasis. This phenomenon is called 'Warburg Effect' [69] and often occurred in certain tumor types such as brain cancer, liver cancer and pancreatic cancer. PGK1 is an important enzyme in the metabolic pathways. Recent studies have revealed that PGK1 can promote cell proliferation and tumorigenesis by enhancing Warburg effect. For instance, Li et. al.'s study reveals that PGK1 functions as a protein kinase to phosphorylate PDHK1, which further promotes the Warburg effect in brain tumorigenesis [101]. Hu et. al recently reported that acetylation of PGK1 can promote cell proliferation and tumorigenesis in liver cancer via glycolysis pathways[102]. Xie et. al.'s study has pointed out that PGK1 is highly involved in MYC-induced metabolic reprogramming, which further causes a reinforced Warburg effect [103]. From the pathway analysis result from section 4.3.2, another significantly enriched 'cellular metabolic process' pathway is observed. This implies the activated Warburg effect in PDAC samples. So far, there are studies that focus on targeting Warburg effect to treat pancreatic cancers. Rajeshkumar et. al has selected a small molecule called 'FX11', which inhibits a lactate dehydrogenase-A (LDH-A), a critical enzyme in metabolizing pyruvate, to block 'Warburg effect' [104]. They observed that for TP53 mutant cells, their approach can significantly increase tumor cell apoptosis. These studies provide the possibilities of targeting 'Warburg effect' to treat PDAC.

Hence, together with the survival analysis result shown in Figure. 11 (c), the findings suggested that PGK1 is a potential target that alternatively aims Warburg effects and thus worth further experiment validation.

‘DNA polymerase epsilon 2’ (POLE2) and ‘Hyaluronan mediated motility receptor’ (HMMR) have previously reported as significantly hyper-expressed in both PDAC tissues and cell-line expression profiles [105]. Studies have linked HMMR and its product ‘Receptor for Hyaluronan Mediated Motility’ (RHAMM) to a variety of hematological malignancies and other solid tumors [106-108]. This is because RHAMM working in concert with BRCA1 and BRAC2, can significantly promote tumor growth and metastasis for pancreatic cancer [109] in vivo, and multiple other cancer types such as basal-like breast cancer [110] and glioma [111] in vivo. Hence, Willemen et. al pointed out of HMMR/RHAMM being a considerable potential target for cancer immunotherapy [112]. Moreover, Li, Ji and Wang has targeted HMMR via long noncoding RNA (lncRNA) and successfully suppressed Glioblastoma in mouse xenograft model [113]. This evidence suggest HMMR and its product RHAMM is worth further study for its potential to be used as PDAC drug target. POLE2 is highly involved in DNA repair and replication. However, targeting POLE2 to treat cancer is rarely reported. Li et. al used β-elemene, which is a type of elemane sesquiterpenoids, to suppress POLE2 expression and restrained lung adenocarcinoma cell malignant in vitro [114], which can be used as a evidence of treating pancreatic adenocarcinoma (PDAC) by targeting POLE2.

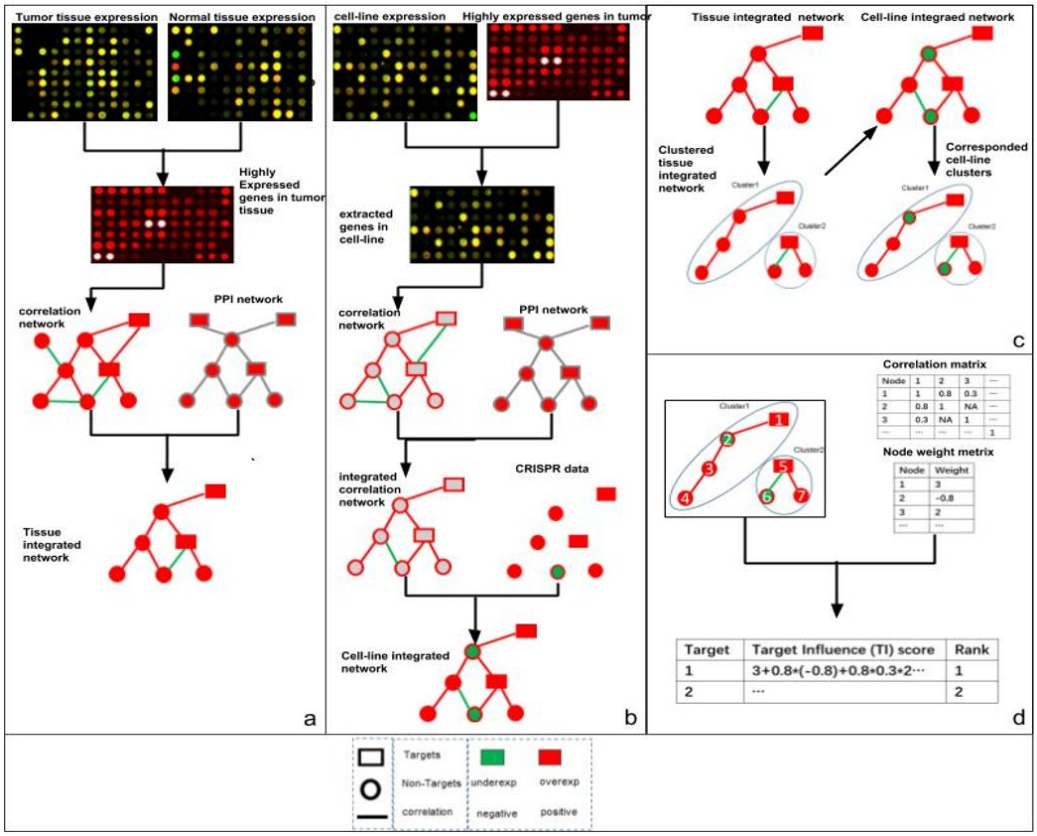

Figure 9 Workflow of 'SCN' algorithm

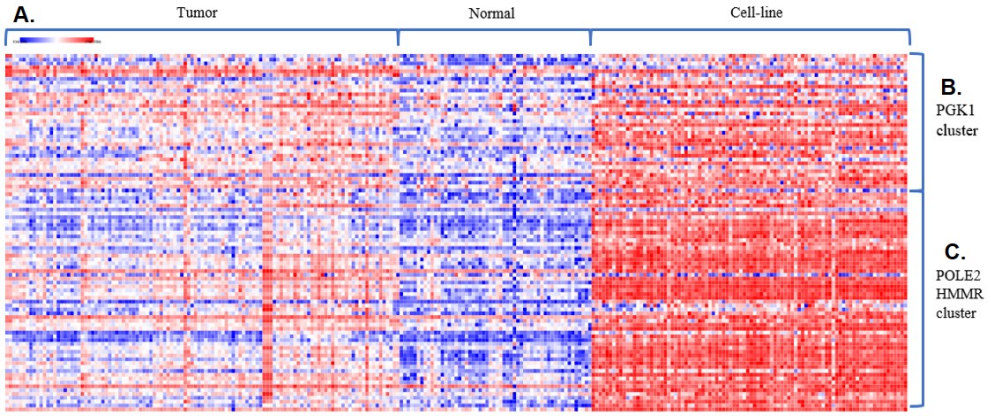

Figure 10 Heatmap of PGK1 and POLE2-HMMR clusters in three groups

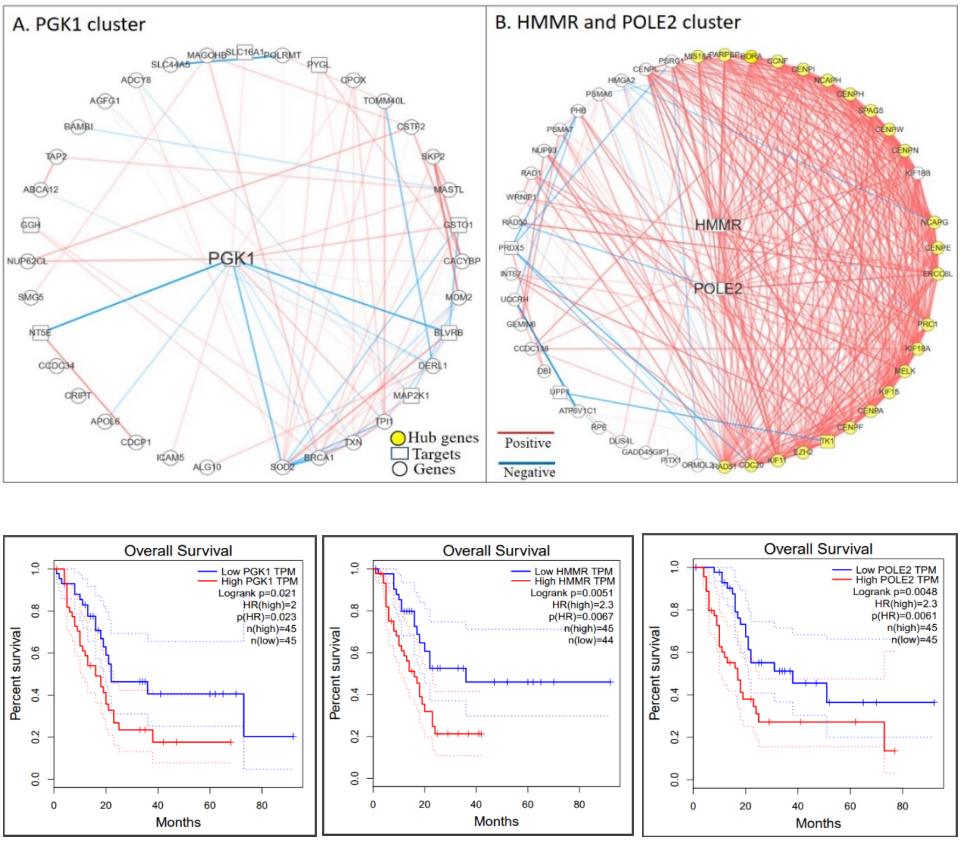

Figure 11 Network graphs and survival curves of top three drug targets

Table 5 Gene expression data used in ‘SCN’ study

NO	Human pancreatic cancer cell line	Human PDAC tumors	Human normal pancreas tissues
1	GSE36133 (43)	GSE42952 (33)	GSE46385 (3)
2	GSE46385 (7)	GSE51978 (2)	GSE16515 (16)
3	GSE21654 (22)	GSE16515 (36)	GSE15471 (39)
4	GSE17891 (20)	GSE15471 (39)	
5		GSE23952 (3)	
Sample size	92	113	58

Table 6 The top 12 ranked drug targets and associated gene expression variation in tumors

Name	RANK	Cancer drug target (Y/N)	Tumor gene expression (Average)	Normal gene expression (Average)	T_v_N log FC	T-value	p-value	Gene essentiality in CRISPR
PGK1	1	N	10.18	9.28	0.90	8.03	<0.01	-1.84
POLE2	1	Y	5.87	4.83	1.04	5.31	<0.01	-1.31
HMMR	2	N	6.83	5.06	1.77	4.31	<0.01	-0.96
VDAC1	4	N	9.53	8.83	0.70	6.30	<0.01	-1.85
PPP2CA	5	N	8.61	8.40	0.21	3.98	<0.01	-1.94
DARS2	6	N	5.63	5.16	0.47	3.02	<0.01	-0.54
TK1	7	N	6.56	5.95	0.61	3.37	<0.01	-0.42
VARS	8	N	5.52	5.14	0.38	3.01	<0.01	-2.13
DHFR	9	Y	7.09	6.45	0.64	3.75	<0.01	-1.06
MMP14	10	N	7.40	6.51	0.89	4.37	<0.01	-0.21
ERBB2	13	Y	6.65	5.58	1.07	3.23	0.01	-0.20
MTOR	32	Y	5.45	5.04	0.44	3.74	<0.01	-1.24

5. SCN Website: Graphical Computation for Prioritization of Cancer Therapeutic Targets Using CRISPR–Cas9 Screen

5.1 Introduction

CRISPR screening technology can estimate the essentiality of given genes under different conditions. Thus, it has more often been applied in selecting novel target for guiding patient treatment. Numerous tools and databases are established to estimate gene essentiality and predict novel targets using whole genome CRISPR screening data across various cell-lines and cancer types. For instance, However, there is still in lack of interactive platforms for users to have customized targets recommendations. Li, Xu and Xiao et al. developed a mean-variance model based analysis tool (MAGeCK) that can prioritize novel essential genes and pathways under different experiment conditions [115]. Another tool called 'NetICS'generated a graph diffusion-based model for prioritizing cancer genes using multiple types of omics data [116]. However, most of these tools can only be run locally without querying other databases, which result in limited resources of omics data for accurately prioritizing targets. Database wise, Rauscher et al. constructed a database 'GenomeCRISPR'for high-throughput CRISPR-Cas9 screenings [117]. 'Depmap'is another CRISPR-Cas9 screening database that covers more than 20 major cancer types and over 400 cancer cell-lines [118]. Behan et al. utilized large-scale CRISPR-Cas9 screening and regression models to differentiate disease specific essential genes, universally essential genes and non-essential genes [119]. These databases don't have a server-client interactive interface. Users cannot upload their own omics data to get a

customized target prediction result. Based on the issues and needs mentioned above, the ‘SCN website’, an interactive tool for delivering customized drug targets and visualizations is developed. SCN algorithm provides a systematic way of prioritizeing potential targets from multi-omics data. And SCN website fills the gap between algorithms and databases, enabling users to upload their own expression data, calculate the result while query the databases.

5.2 Materials and Methods

5.2.1 Data content

After crossing match data from all databases, we include and provide selected targets for 10 cancer types, namely: Colon adenocarcinoma (COAD), Esophageal carcinoma, (ESCA), Head and Neck squamous cell carcinoma (HNSC), Kidney renal papillary cell carcinoma (KIRc), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Pancreatic adenocarcinoma (PAAD), Thyroid carcinoma (THCA). These 10 cancer types have tissue expression data, corresponding cell-line data and CRISPR gene essentiality data and thus been included.

Cell-line data across various cancer types cell-line expression data and CRISPR-Cas9 screening data (v3.3.8) are downloaded from database ‘Depmap’[118, 120, 121] (https://portals.broadinstitute.org/achilles). These whole genome data is across >400 cell-lines and 24 most common cancer types.

Protein-protein interaction network A protein-protein-interaction network consisting of 9,606 proteins (version ‘9606.protein.links.v11.0’) from STRING

database [32] is included in this study. It is a comprehensive database (https://string-db.org/) that accumulates prior knowledge of biological pathways and protein-protein interactions.

FDA approved drug targets all FDA approved drugs from Drug bank [122] (version 02/03/2018) are downloaded along with their targets. These targets are defined as FDA approved drug-targets. In total, there are 1,317 targets included in the database, which contains 283 targets of cancer drugs.

Example tissue expression data of pancreatic cancer a set of example pancreatic tissue expression data from Gene Expression Omnibus [61] (GEO, http://www.ncbi.nlm.nih.gov/geo/) database is provided. This set consists of 113 Pancreatic ductal adenocarcinoma (PDAC) tissue samples and 58 adjacent normal pancreas tissue samples, all of which are generated from Affymetrix Human Genome U133 Plus 2.0 Array, which contains 54,675 probes pointing to over 20,000 genes.

5.3 Implementation

SCN is a free and publicly available website (https://pcm2019.shinyapps.io/SCNrank/). Both server and client ends are carried out by 'Shiny' package, which is a convenient solution that integrates PHP and back-end databases for setting up an interactive website on an R server. The website consists of four tiers (Figure 12): Data tier stores preprocessed omics data for the analysis; Computation tier executes SCN algorithm with data from both data tier and client to generate results. Interface tier allows users to upload their inputs, reflect the progress of the algorithm execution and provides sample input/ouput files; Output tier constructs

a table of statistics of ranked targets. Top targets are displayed in network images. Kaplan-meier curves are drawn for top targets to indicate how hazardous they are for patient survival.

All omics data is stored locally as inputs of the algorithm. The SCN algorithm is written in R (3.6) and python (2.7.16), in which R program does the clustering and python program does the scoring. Networkx (1.9) package is called in python program for representing subnetworks and searching paths among nodes in subnetworks. Results of the algorithm are displayed as tables, Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG) images. Tables are generated by R, network images are generated by Cytoscape (3.3) and Kaplan-meier curves are called from GEPIA database [123] using TCGA patient annotations.

5.4 Functionalities

Figure 13 generally lists all available results generated from SCN website. Subgraph A indicates workflow of SCN algorithm. Red, green, and yellow blocks indicate input from database, input from user and output respectively. Subgraph B shows an example output of ranked targets. Subgraph C displays an example of Kaplan-Meier curve from survival analysis. Subgraph D indicates an example output from Pathway analysis. Subgraph E indicates an example of corresponding clinical trials.

Survival analysis

SCN website queries GEPIA, which stores all TCGA and GTEx patient annotations, to get survival analysis result. The Kaplan-Meier curves, cox proportional

hazard ratio and the 95% confidence interval information are included in the plot (Figure 13 (c)). Clicking each gene in the list results in a Kaplan-Meier curve, which measures the survival differences of patients with high expression versus low expression levels of this gene.

Clinical trials Proving evidence of clinical trials can connect prioritized targets with clinical research/studies. SCN website provides clinical trials across six cancer types: pancreatic cancer; breast cancer; osteosarcoma; soft tissue sarcoma; ovarian cancer and prostate cancer. As indicated in Figure 13 (B), these clinical trials information includes case number, treatment plan, overall survival and disease-free survival (DFS). All clinical trials information is gathered from Clinical Trials database from NIH (https://clinicaltrials.gov/).

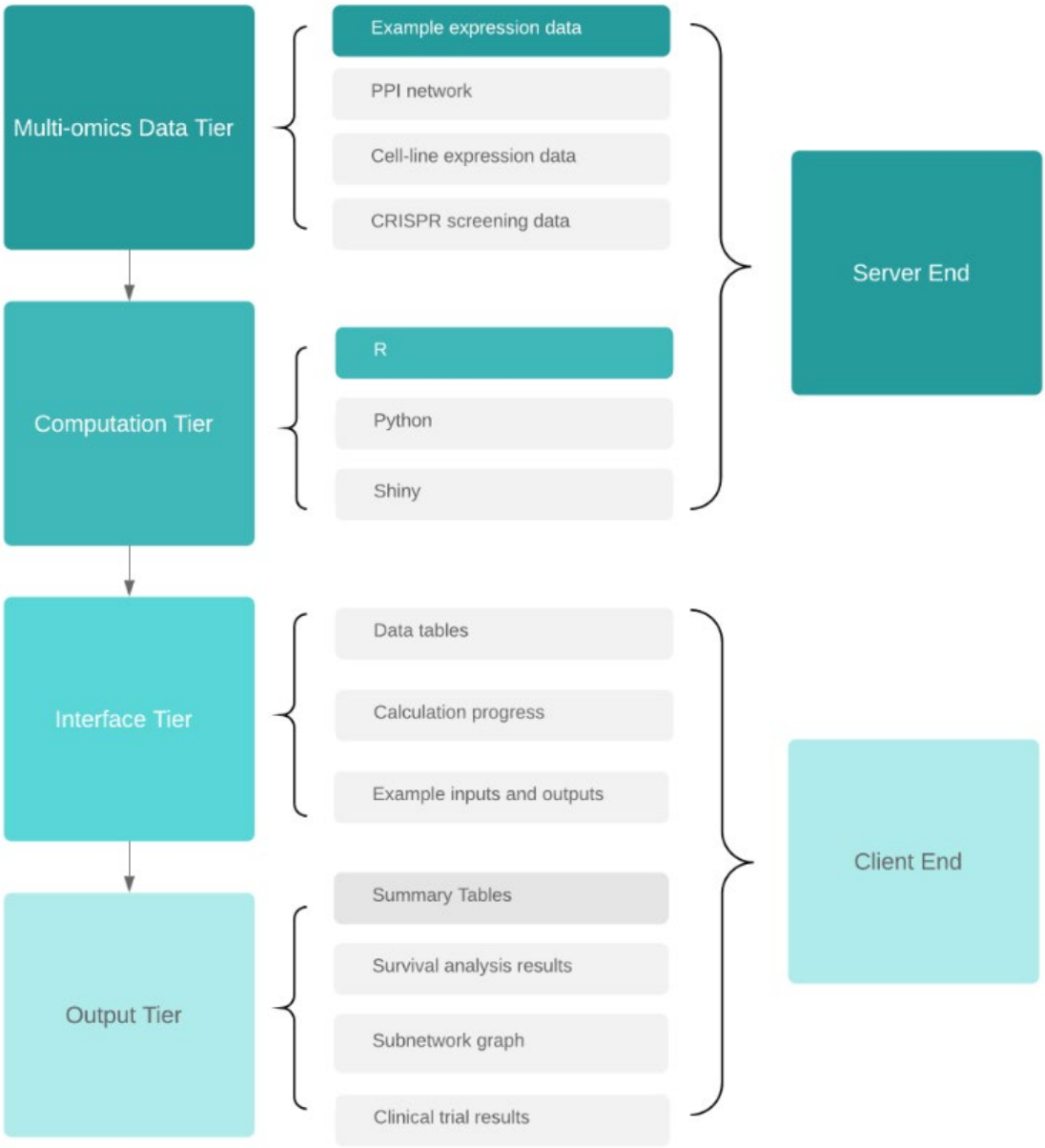

Figure 12 Website Structures of data processing, output type and visualization

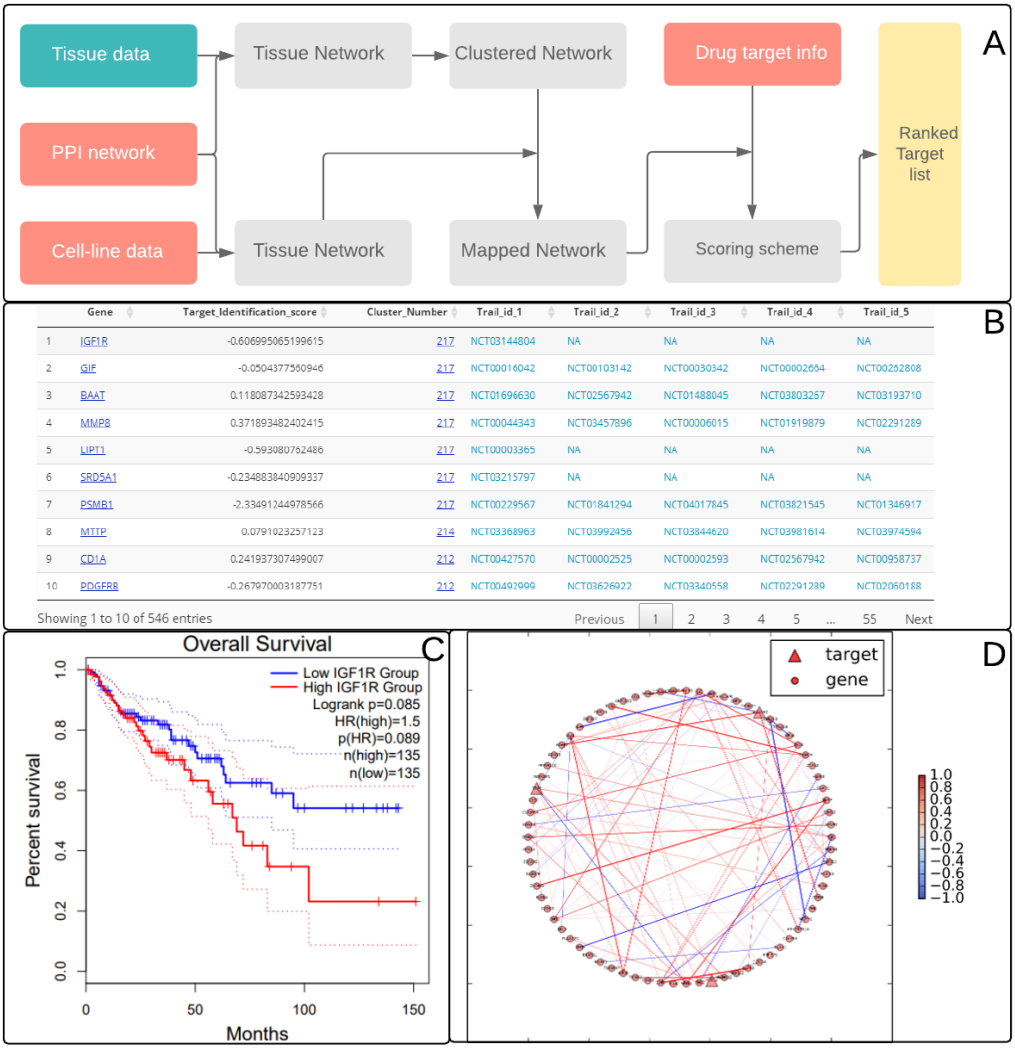

Figure 13 Example outputs of SCN website

6. D-SCN: A Spectral Clustering based Network approaches for Double drug-Targets Prioritization for Cancers

6.1 Introduction

Pancreatic cancer is the third most lethal cancer types in the united states [124]. The American Cancer society has estimated that over 56,000 new cases and over 45,000 new deaths of pancreatic cancer will occur during 2019 [124]. Over 85% of pancreatic cancer types are pancreatic ductal adenocarcinoma (PDAC). Compared to all other major cancer types, PDAC has the poorest 5-year survival (~8%). This is mainly due to late diagnosis: patients are usually diagnosed with PDAC at very late stages when tumor cells have already spread and metastasized to other organs, which limits the efficient treatment plans [125]. Moreover, Due to the treatment-recalcitrant characteristics [126, 127], PDAC cells often develops drug resistance, which undermines the effectiveness of both chemo drugs and targeted drugs [128]. Combinational drug treatment is considered as an alternative approach for its synergistic anticancer effects and reduced resistance of cancer cells [129]. Combinational therapy such as Gemcitabine plus Erlotinib or Gemcitabine plus Nab-paclitaxel has been used in clinic for treating PDAC, however, the survival is modestly improved [126]. This creates an unmet need of identifying novel combinational therapies.

Cell-lines derived from human cancers are widely used for understanding cancer biology and for testing hypotheses to facilitate the cancer treatment [130]. Despite that cell-line models allow investigators to observe and mimic a complete

PDAC process, the complexity of modelling PDAC cell-line has increased the difficulty [131]. Recent advances of CRISPR-Cas9 genome editing technology [132] enables large-scale preclinical testing and validation of therapeutic targets discovery [133]. Protein-coding genes within the whole genome can be simultaneously screened for the evaluation of their essentiality towards cancer cells [24, 134]. In the meantime, genome/transcriptome data generated from high-throughput technology offers a wealth of information for revealing the molecular mechanisms of PDAC, including cancer related genes that can potentially be targets [135]. These different types of omics data interpret PDAC from different angles. However, constructing models that utilize both CRISPR-Cas9 screening data and transcriptome date collected from patients and cancer cells to better prioritize potential targets remains to be explored.

So far, various attempts of using omics-data to identify drug targets have been reported, including those either solely using one types of omics data or integrating multiple types of omics data.

Single drug target selection has been extensively studied recently while various methods have been developed. For instance, ‘Connectivity map’ project (C-map) curated expression profiles of human cells exposed to thousands of drugs, which can be served for drug repositioning [56]. Ma et. al developed an algorithm named ‘Met-express’ [57] that combines gene co-expression network with human metabolic network to predict drug targets for pancreatic cancer. However, these methods only utilize expression data as fundamental knowledge and incorporate other biological knowledge to predict targets. Moreover, Expression level regulation might not

eventually reflect on protein level regulation, on which most drugs function. Additionally, their analysis lacks the support of cell survival phenotypes that directly reflect the effect of gene knockdown/knockout experiment.

Current methods for identifying combinational targets generally rely on network controllability theory [136]. Candidate target combinations are selected based on their controllability and potential synergistic effects over the networks. For instance, 'OptiCon' algorithm [49] takes advantage of existing pathways and mutation information to generate functional networks from expression data. Functional networks are further divided into subnetworks using bipartite network simplification approaches. Optimal control node for each network is identified and synergistic scores among optimal control nodes are calculated. 'VIPER' algorithm [58] applies the Master Regulator Inference algorithm for identifying core regulators from a gene regulatory network generated from gene expression data. These synergistic core regulators can potentially be considered as candidate combinational targets. These methods focus on investigating the synergistic effect among proteins, which can be treated as potential combinational targets. However, neither methods consider the durability of these protein combinations. RACS [137] and DrugComboRanker [48] combines drug-treated expression profiles with pathway information and anticancer drug information to prioritize combinational drugs from existing drugs from constructed functional networks. Taken together, all these methods are conducted mainly on expression level. Other genomic properties such as gene essentiality derived from CRISPR-Cas9

screenings are not evaluated or estimated, which might result in a failure of applying these combinations *in vitro* or in clinic.

To resolve the stated issue, a method called ‘DSCN’ is proposed. It combines expression from patients and cell-lines with CRISPR-Cas9 data and proteomics data, to prioritize combinational targets. Inherited from SCN, a method that combines the same omics data to prioritize single target, DSCN constructs functional networks and mimics expression and essentiality changes of genes after targets being sequentially knocked-out. Combinational targets are then selected based on the sequential process. Moreover, another method called ‘DSCNi’, which utilizes similar processes of DSCN for predicting combinational targets for individuals, are also proposed. DSCNi takes one patient’s expression profile as input to predict ranked combinational targets for various cell-lines.

6.2 Materials and Methods

6.2.1 Materials

Expression data (cell-line and tissue), perturbation data, PPI network and Drug target information are included as multiple types of omics data in the study. The composition and version of each omics-data type can be found from Table 7.

Expression profiles We mined the entire Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database and identified 263 samples in total, including 92 PDAC cell-line samples, 113 PDAC tissue samples and 58 adjacent normal pancreas tissue samples (Table 7). All these samples are generated from Affymetrix Human Genome U133 Plus 2.0 Array, which contains 54,675 probes

pointing to over 20,000 genes. Even though there are other expression samples from other numerous platforms, samples only from U133 Plus 2.0 Array are kept for further analysis. This is because it is the most common platform of PDAC samples in GEO; it covers the entire genome; data generated from this platform can be effectively integrated. Notably, all 92 cancer cell-line samples mined from GEO are also part of 'Cancer Cell Line Encyclopedia' (CCLE) database [138], which offers a wealth of transcriptome profiles for more than 500 cell-lines across 30 different cancer types.

In additional, for benchmark with other approaches, which requires gene mutation information, 183 PDAC tissue samples from 'The Cancer Genome Atlas' (TCGA) database, which provides both expression profiles and mutation information of tissue samples, are included in the study.

Protein-Protein Interaction network STRING [19] is a comprehensive pathway database (https://string-db.org/), which accumulates numerous biological pathways and protein-protein interactions (PPI) derived from experiments and computational inference. PPI network version 11 data is included in the study from STRING.

Gene essentiality from CRISPR-Cas9 screening data 'Project Achilles' [80-82] (https://portals.broadinstitute.org/achilles) database offers a broad range of gene essentiality data derived from various perturbation experiments, including RNAi and CRISRP-Cas9. We choose CRISPR-Cas9 over RNAi for its less off-target effects [83], which is critical for accurately selecting drug targets for cancer. In total 22 PDAC cell-lines that consists of 74,222 sgRNAs are included in this study.

Existing drug targets DrugBank [139] database contains all FDA approved drugs and corresponding targets. All these drugs are downloaded and then mapped them onto 1,769 targets, of which 313 genes are targets of cancer drugs.

Existing synthetic lethal pairs Synthetic lethality (SL) reflects a type of genetic interaction between two genes such that simultaneous perturbations of the two genes result in significantly more cell death than cell death caused by perturbating either one of genes [140]. SynlethDB database consists of SL pairs curated from biochemical experiment, other related databases, computational predictions and text mining [140]. Documented SL pairs and their SL scores from SynlethDB are utilized for benchmarks and validations.

Drug sensitivity data NCI60 database [141] offers numerous drug sensitivity measures over a broad range of cancer cell-lines. drug sensitivity information from NCI60 are included to validate predictions from DSCNi.

6.2.2 Methods

Defining the influence of targets towards the network Given an undirected network $G =< V, E >$, where V is the node set and E is the edge set. A target node T can cause towards after G being knocked-out can be defined as the sum of its influence towards every other node:

$$TI(T) = \sum_{i \epsilon V} s[N_i | Pa(N_i)] \quad (28)$$

Where TI indicates target information, N indicates a node, $Pa(N_i)$ indicates the parent set of N. The term $s[N|Pa(N)]$ thus indicates the influence towards N.

Hence, the objective is to maximize formula (28).

Different routing methods will cause different $Pa(N)$, which further causes different s. In this study three different routing methods, which are displayed in Figure14, are examined and benchmarked.

Assume that edge weight indicates the correlations between nodes. Consider T_1 the target to be knocked out, the influence it can cause with three different routing methods can be defined as:

Heaviest (Most correlated) Paths: in an undirected network, the distance between two nodes are defined as the distance of shortest/most correlated path. In Figure.14(b), each node N has one heaviest path to T_1. In this case, the parent set $Pa(N) = V$, where V indicates the set of nodes in the heaviest path.

Random Walk (Figure 14(c)): starting from T_1, each node N is randomly visited. This process takes certain number of $Pa(N) = N_i$ where N_i is the first node that visited N.

Diffusion: starting from T_1, each node is visited in a hierarchical order. $Pa(N) = V$, where V indicates the parent nodes from the higher tier and connected sibling nodes from the same tier. For instance, in Figure14(d), when T_1 is knocked out, a tree-like hierarchical structure is constructed using T1 as root to point out the tiers being sequentially influenced. Clearly there are 5 tiers in the hierarchical structure and the influence is transmitted from tier 1 to tier 5. The rationale behind it contains two folds: 1. Since we are handling an undirected network, a parent set is defined as a set of its directed parents and connected sibling nodes. Biological network is a complex and redundant system. 2. Influence can echo between connected sibling nodes.

Eventually diffusion method is selected and applied in the DSCN and DSCNi algorithm. The evidence of choosing diffusion method among three routing method in is described in section 6.3.1.

As indicated in Figure 15, the algorithm generally consists of five steps:

Step 1: Construct functional networks. In this step two integrated functional networks, namely tissue network G_t and cell-line network G_c are constructed.

G_t consists of a skeleton from STRING PPI network, edge weights from pair-wise Pearson correlations between genes in tumor samples, and node weights from tumor versus normal fold change. Assume there are in total n genes (nodes) in G_t, a affinity matrix S_t is constructed to denote the edge weights:

$$S_t = \begin{pmatrix} 0 & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & 0 \end{pmatrix} \quad (29)$$

Where $w_{\mathrm{ab}}, a, b \in (1, n)$ in the matrix indicate the edge weight (correlation) between gene a and b in the tissue network.

A diagonal matrix is defined to indicate the node weights

$$D_t = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix} \quad (30)$$

Where w_i in the matrix is the tumor versus normal fold change of gene i.

G_c consists of an identical skeleton from the same STRING PPI network, edge weight from pair-wise gene correlations in cell-line samples, and node weight from CRISPR-Cas9 screening data. Similarly, we defined affinity matrix S_c and diagonal matrix D_c to indicate the edge weights and node weights in cell-line network G_c.

Step 2: Construct corresponding Laplacian matrices. Laplacian matrices are then constructed for representing the network structure of G_t and G_c. The Laplacian matrix is defined as:

$$L = D - S \quad (31)$$

Where D and S indicate diagonal matrix and affinity matrix, respectively. L_t and L_c are constructed for tissue network and cell-line network, respectively.

Step 3: Perform spectral clustering to tissue Laplacian matrix. To obtain disease related subnetworks, a spectral clustering is performed on G_t, which can be generally described as: for the Laplacian matrix L_t, a eigen-decomposition is performed while eigenvectors that corresponds to k smallest non-null eigenvalues are subsequently extracted. K-means algorithm is then performed for these k eigenvectors to get clusters (subnetworks). The steps of performing spectral clustering can be described as:

Normalize Laplacian matrix L_t to L'_t

$$L'_t = \begin{pmatrix} abs(w_1) & \cdots & -\frac{abs(w_{1n}w_1)}{\sum_{k=1}^{n} abs(w_{1k})} \\ \vdots & \ddots & \vdots \\ -\frac{abs(w_{n1}w_1)}{\sum_{k=1}^{n} abs(w_{nk})} & \cdots & abs(w_n) \end{pmatrix} \quad (32)$$

In this step, L_t is normalized to L'_t so that L'_t becomes positive semi-definite. Here $abs()$ indicates the absolute operation. In Laplacian matrix L'_t, all diagonal elements are positive while all other elements are negative. And row sum of non-diagonal elements is equal to its corresponding diagonal.

Perform eigen decomposition for matrix L' and obtain the spectrum E=$\{\lambda_1, \lambda_2 \ldots \lambda_n\}$, where λ_i indicate the ith smallest eigen values. Decomposition for a

positive semi-definite matrix results in a set of non-negative eigen values: $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Pick k smallest non-negative eigenvalues $\{\lambda_i, \ldots, \lambda_{i+k}\}$ and their corresponding eigenvectors, each of which is a $\mathrm{n} * 1$ vector. Combining k eigenvectors results in a $n * k$ matrix.

Now the combination results in $n * k$ matrix, where n corresponds to n nodes and each of them has k values (coordinates) in this matrix. These values can be used to calculate pair-wise Euclidean distances. K-means clustering is then performed for n nodes. The initial number of K' in K-means is not associated with number of selected eigenvectors k. To select the K' that result in a good fitness, an intuitive approach is applied: from K'=2 to the number of total number of variables, K-means algorithm is continuously executed while Hartigan's number is calculated at the same time. Hartigan's number is a measure of the clustering quality by comparing two clustering results. The first K' that result in Hartigan's number less than 10 is then selected. Obviously, this scheme of picking K' doesn't guarantee global optima.

K-means clustering leads to K' exclusive clusters (subnetworks), each of which contains several node names. From the whole network G_t, subnetworks are then extracted according for each cluster. The corresponding subnetworks of these clusters are named $g_{t_1}, \ldots g_{t_k}$ respectively.

Step 4: Map tissue subnetworks onto cell-line subnetworks and measure the similarities. Since G_t and G_c shares identical network structure, for each $g_{t_1}, \ldots g_{t_k}$, corresponding subnetworks $g_{c_1}, \ldots g_{c_k}$ are extracted from cell-line network using their

common node names. Till this point, clustering results from Tissue subnetworks have been mapped onto cell-line subnetworks. For each subnetwork pair <g_{t_i}, g_{c_i}>, $i \in (1, k)$, the similarity between tissue and cell-line are measured using the following scheme.

For each <g_{t_i}, g_{c_i}>, $i \in (1, k)$, to make them comparable, cell-line diagonal matrix D_{c_i} is firstly normalized to tissue diagonal matrix D_{t_i} using the following formula:

$$D'_{c_i} = \begin{pmatrix} \frac{w_{c,1}\sum_1^m w_{t,k}}{\sum_1^n w_{c,k}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{w_{c,1}\sum_1^m w_{t,k}}{\sum_1^n w_{c,k}} \end{pmatrix} \quad (33)$$

Where $w_{c,k}$ and $w_{t,k}$ denote weight of node $k \in (1, m)$ in cell-line network and tissue network respectively. m is the total number of nodes in g_{t_i} and g_{c_i}. Till this stage, the node weight is normalized.

For each <g_{t_i}, g_{c_i}>, $i \in (1, k)$, two Laplacian matrices are constructed using formula (34) and (35).

$$L_{t_i} = D_{t_i} - S_{t_i} \quad (34)$$

$$L_{c_i} = D'_{c_i} - S_{c_i} \quad (35)$$

Obviously, after node weight normalization, Trace (L_{c_i})=Trace (L_{t_i}). Same traces indicate same spectrum (eigenvalues), which further indicate two matrices are in the same scale. Now edge weights (non-diagonal elements) are adjusted accordingly using the following formula:

$$L'' = \begin{pmatrix} w_1 & \cdots & \frac{w_{1m}abs(w_1)}{\sum_{k=1}^{m} abs(w_{1k})} \\ \vdots & \ddots & \vdots \\ \frac{w_{m1}abs(w_1)}{\sum_{k=1}^{m} abs(w_{mk})} & \cdots & w_m \end{pmatrix} \quad (36)$$

Where w_i indicate node weight of node i, w_{ik} indicates edge weight between node $i\ and\ k$, $i,k \in (1,m)$. M is the total number of node within each subnetwork. We normalize matrix for both $g_{t_i}\ and\ g_{c_i}$ to get normalized Laplacian L''_{t_i} and L''_{c_i}. Till this step, all edges (non-diagonal elements) in both Laplacian matrices acquired node features during normalization. Notably, node weight and edge weight can be either positive or negative in L''. The original directions (positive or negative) of node weights and edge weights are kept in the process to accurately measure their distances.

For two corresponding subnetworks $g_{t_i}\ and\ g_{c_i}$, their distances between their Laplacian L''_{t_i} and L''_{c_i}, can be calculated using the following schema:

$$D(g_{t_i}, g_{c_i}) = \sum_{i=1}^{m} \sum_{j=1}^{m} (L''_{t_i}(i,j) - L''_{c_i}(i,j))^2,\ i \neq j \quad (37)$$

The term $L''(i,j)\ \ i \neq j$ indicate the edge weight between node i and j in a given Laplacian matrix. The term $(L''_{t_i}(i,j) - L''_{c_i}(i,j))^2$ indicate the Euclidean distance between the same edge in two Laplacian matrices.

Step 5: Score and rank target combinations with cell-line subnetworks. This step can be further divided into three processes: score first target, adjust network, score second target given first target. Here this score is defined as Target influence (TI).

For each target T_a of FDA approved drugs, the diffusion routing method is applied for measuring its influence towards other nodes. In Tier 1, T_1 is the target being knocked out, it does not depend on any other node but itself. Hence, the

influence $S(T_1)$= W_{T_1}, where W_{T_1} is the node weight of T_1. In the following tiers, for each node, the influence it can cause consists of two parts: the influence from itself and the transmitted influence from its parent set. For instance, in Fig. 14(d), the influence of node N_5 can be described as: $S(N_5) = W_{(N_6,N_5)}W_{N_5} + W_{(N_3,N_5)}W_{N_5}$. In summary, the total influence that a network can get after a target T_{a} being knocked out is the sum of the influence of all nodes can get within the network, which can be further described as:

$$TI(T_a) = S(T_{\mathrm{a}}) + \sum_{k\in g_{c(T_a)}} S(N_{\mathrm{k}}) \quad (38)$$

Where T_{a} is the target and N_{k} represents any other member genes in this subnetwork $g_{c(T_a)}$. In other words,

$$TI(T_a)=W_{T_a} + \sum_{k\in g_{c(T_a)}} \sum_{N_j\in P_{N_k}^{g_{c(T_a)}}} W_{N_k} W_{(N_k,N_j)} \quad (39)$$

Where m is the total number of member genes within the subnetworks $g_{c(T_a)}$. $P_{N_k}^{g_{c(T_a)}}$ is the parent set of N_k with the subnetwork $g_{c(T_a)}$. N_j is any member node within the parent set. W_{N_k} indicates the node weight of N_k. $W_{(N_k,N_j)}$ indicates the edge weight between N_k and N_j.

After one target being knocked out, the corresponding subnetwork will change in terms of node weight and edge weight. Due to the lack of expression data of before/after CRISPR-Cas9 screening, these changes need to be inferred. For a target being hypothetically knocked-out, its expression level varies across samples. We divided samples into two groups based on its mean expression values and the low expression group is used to represent its expression state after being knocked-out. Hence, the new pairwise correlation of other genes are generated from the low

expression group and updated to the network as edge weights. Similarly, the essentiality score is recalculated using the less essentiality subgroup of the given target. Till this point, the node weight and edge weight are updated in the corresponding subnetwork. In DSCNi, since expression profile (both tissue and cell-line) of one sample is taken as input, only edge weight will be changed in the corresponding subnetwork.

Given a target T_{a} being knocked out and an adjusted subnetwork, for each of the rest target T_{b} of FDA approved drugs, the combined influence of $<T_a, T_b>$ is measured with the following formula:

$$TI(T_a, T_b) = TI(T_a) + TI(T_b|T_a) \quad (40)$$

Where $TI(T_a)$ is defined in formula 11 and $TI(T_b|T_a)$ is defined as:

$$TI(T_b|T_a) = W_{T_b}^{g_{c(T_b|T_a)}} + \sum_{k' \in g_{c(T_b)}} \sum_{N_{j'} \in P_{N_{k'}}^{g_{c(T_b|Ta)}}} W_{N_{k'}} W_{(N_{k'}, N_{j'})} \quad (41)$$

Where $g_{c(T_b|T_a)}$ indicates the subnetwork $g_{c(T_b)}$ after T_a being knocked out. The weight of the network may or may not be adjusted by this operation. If two targets are not in the same subnetwork, then $g_{c(T_b|T_a)} = g_{c(T_b)}$, which indicates subnetwork $g_{c(T_b)}$ are not adjusted based on the knock-out of T_a. If two targets are in the same subnetwork, then $g_{c(T_b|T_a)} \neq g_{c(T_b)}$.

The first term indicates the influence of second target T_b towards its subnetwork $g_{c(T_b)}$, given the fact that its weight might or might not be adjusted after T_a being knocked out. The second term indicates influence from nodes not in tier 1.

6.3 Results

6.3.1 Routing method selection

In section 6.2.1, three different routing methods are proposed for evaluating the influence of a given target towards the network. These methods are benchmarked by calculating the score difference between synthetic lethal pairs and non-synthetic lethal pairs. 23 Pancreatic specific SL pairs exist in SynlethDB. These SL pairs correspond to 15 genes. Two groups are then constructed: a group of 46 SL pairs (23 pairs with order) and a group of 164 non-SL pairs, which are all possible combinations of 15 genes minus 46 SL pairs. Even though SL pairs are not drug targets, and DSCN only calculated combination scores for existing drug targets. We assumed them to be drug targets and calculated their combination scores for three different methods. Independent T-test was performed between SL and non-SL pair groups. Two scales of network are also treated as variables in the benchmark: ‘global’ network indicates whole genome network before spectral clustering process mentioned in section 6.2.2 step 3; ‘local’ network indicates subnetworks after spectral clustering process. Targets influence different number of nodes within two different networks.

From Figure 16, it can be observed that SL pair group has a significantly lower score than non-SL pair group while using diffusion method under local network. This segregation fails to occur with other routing methods under both local and global network. This proves the necessity of performing spectral clustering and choosing diffusion method in DSCN.

6.3.2 Benchmark between DSCN, VIPER and Opticon

To measure the accuracy of DSCN, we compared its predictions with predictions from two other algorithms: Opticon and VIPER, both of which use expression profiles to predict synergistic master regulators that control the entire subnetworks. VIPER takes expression profile from both tumor and normal samples to predict the corresponding regulon, which is a set of regulators and their targets. On top of that, it infers synergies among regulators. Opticon requires both expression profile from tumor samples and corresponding mutation information as input to infer first infer master regulators then predict synergies among them. PDAC expression data described in Table 7 is used and all 6,550 PDAC mutations obtained from TCGA as common input to benchmark the predictions for three algorithms.

In total, DSCN, VIPER and Opticon predicted 37,275, 2778 and 191 combinations. All of them are mapped onto SynlethDB, which consists of 12,821 SL pairs. Surprisingly, neither VIPER nor Option has any overlap with documented SL pairs while DSCN has 936 overlapped synergistic combinations. We then compared the rank of 936 overlapped SL pairs and 79 overlapped PDAC specific SL pairs using spearman correlation test (Table 8). Under both PDAC specific and all documented SL pairs settings, predicted ranks from DSCN maintains a relatively high correlations with documented SL pairs in SynlethDB. Moreover, predicted scores of 79 PDAC specific SL pairs are significantly lower than scores of 6,162 random combinations under independent T-test ($T=-1.85$, $p=0.05$). This shows consistency of T-test result from section 6.3.1.

All benchmark work has been done on supercomputer ‘Carbonate’ from Indiana University. DSCN spent 12 hours on a single CPU core to complete the search. shortest runtime among three algorithms. Although DSCN provided the option of executing algorithms in parallel, in the benchmark it is executed with single thread. Most of the computational time of VIPER is on generating a whole genome mutual information network using ARACNe [142], which requires ~141 hours to reconstruct such large scale network. Opticon mainly consists of two computational tasks: calculating subnetworks and calculating null distributions, each of which costs ~160 computation hours. Even though Opticon optimize the calculation by simultaneously submitting hundreds of jobs on supercomputer. The overall speed remains the slowest.

6.3.3 Top ranked drug combinations and associated subnetworks

For all overlapped 1,437 drug targets, all possible combinations among them are calculated using PDAC data. Notably, genes involved in top 230 combinations are from the same subnetwork (No.42). Table 9 displays the top 10 ranked target combinations. All of them have relatively close TI scores. Survival analysis were performed for calculating hazard ratio of high expression group versus low expression group using clinical information from TCGA. Interestingly, most of top ranked combinations are TFRC plus another genes and most TFRC involved combinations show significant high hazard ratios. Transferrin receptor (TFRC) encodes a surface receptor for cellular iron intake, It is reported to be highly expressed in PDAC and high associated with PDAC cancer growth and survival [143]. Recent studies suggest several key pathways of inducing Ferroptosis, such as MAPK and Ros [144]. Hence,

targeting upstream genes (e.g. MAP2K2, EGLN2) along with downstream genes (e.g. TFRC, FTL) might lead to a synergistic effect of inducing Ferroptosis. This explains why combinations listed in Table 9 are in the same subnetwork and have TI scores.

Figure 17 and Figure 18 display the tissue specific network and cell-line specific network of subnetwork No.42. In both figures, red and blue shades on edges indicate positive and negative correlations. Red/blue shades on nodes indicate over-expression/under-expression in tissue network and positive/negative gene essentiality in cell-line network. Dimond and round shapes of node indicate target genes and non-target genes, respectively. Shades of colors indicate intensity. Due to the high density of the subnetwork, only edges associated with target nodes are drawn. Edges among target nodes are highlighted with colors. The distance between these two networks are 16.65 (min=0, max=239) and ranked 100 among all 260 subnetworks. Given the fact that this subnetwork is the fourth largest subnetworks among all subnetworks, two networks share significant similarities. These similarities also reflect on the nodes and edges. All target nodes, most of the non-target nodes and edges share the same colors (directions). Together with the identical network topology, cell-line and tissue data shows high consistency on this subnetwork.

6.3.4 Comparison between predictions of DSCNi and existing drug synergies in cell-lines

Due to the limited number of PDAC cell-lines (N=1) and related combinational drug treatments (N=5), predictions of DSCNi using Triple negative breast cancer (TNBC) cell-lines are used and tested. All TNBC cell-lines from various databases are

examined to check the overlap, including transcriptomic data from CCLE database, drug sensitivity data from NCI60 database and CRISPR gene essentiality data from 'Project Achilles'. 'HS578T' is selected as the candidate cell-line and its omics data is extracted as well, including its gene essentiality value and drug sensitivity. Transcriptomic data of 12 TNBC cell-lines from CCLE and 115 tissues from TCGA are selected accordingly. Target combinations for sample 'TCGA-A1-A0SK' via cell-line HS578T are predicted and then validated using drug sensitivity from NCI60 database.

In total, scores for all 2,063,532 combinations of 1,437 existing drug target are predicted. Among them, 14,067 overlaps with 1,031 drug combinations. Bliss score [145] measures expected effect of two compounds being independently acting. Two compounds are considered to be synergistic if their bliss score>0.12 [146]. Firstly bliss scores between 1,031 overlapped drug combinations and all 5,227 drug combinations for 'HS578T' are compared. Bliss scores of overlapped groups are significantly higher than all drug combinations (T=2.35, P=0.02).

A high Drug synergy (bliss score) does not necessarily reflect a high lethality. It reflects the additive effects when two drugs are simultaneously acting. To validate the predictions from DSCNi using drug synergy data, the synergistic effect between target combinations and drug synergy need to be compared. Here we defined a predicted target combination to be synergistic if their TI score is higher than the sum of TI score of each target. A contingency table (Table 10) is created for measuring the association between predicted target combinations and corresponding drug combinations. In the table, PS, PN, DS and DN indicate predicted synergistic target

combination, predicted non-synergistic target combination, target combinations reflected in drug synergies, target combinations not reflected in drug synergies respectively. The phi correlation is defined as:

$$\phi = \sqrt{\frac{\chi^2}{n}},\ \phi \in [-1,1] \quad (42)$$

where χ^2 indicates the chi-square statistic. $\phi = \pm 1$ reflects maximum positive/negative correlations between rows and columns in contingency table.

Predicted synergistic target combinations (PS) has a high positive correlation (0.32) with target combinations reflected in drug synergies. Notably, no PS & DN combinations occurs in Table 10, indicating that predictions from DSCNi has very few false positive rates.

7,069 out of 14,067 discovered target combinations are identified as synergistic combinations. Top 72 of them ranked by TI score are associated with MAPK3 and another genes, which further point to ARSENIC TRIOXIDE and another compounds. This is due to the strong synergistic effect between MAPK3 and many other genes. MAPK3 is often highly expressed in TNBC due to the activation of Ras/MAPK pathway (Hazard ratio =1.5, $p<0.01$). Signals from Ras pathway are transmitted through Raf, MEK, MAPK1/3 to the nucleus to initiate downstream transcription factors such as MYC, ETS-1, ETS-2, and ELK-1, which further regulate cell proliferation and survival [147]. Inhibition of Ras/MAPK pathway has been proven as highly effective in treating advanced melanoma [148] and preventing/treating TNBC metastasis *in vivo* [149, 150]. ARSENIC TRIOXIDE (As_2O_3), which has been successfully applied in treating hematologic malignancies due to its capability of inducing cell apoptosis, has not been

widely applied on treating solid tumors, including TNBC. Recent studies suggest that inhibiting pathways such TNF, JAK and PI3K-Akt work in concert with As_2O_3 for better treating TNBC [151-154]. These findings coincide with the predictions of DSCNi.

Other combinations in Table 11 are selected for their high bliss scores, TI scores and documented synthetic lethality. Among them LAPATINIB, DOCETAXEL, PACLITAXEL are the most common chemo drugs for treating TNBC and (PACLITAXEL, LAPATINIB) combinations are widely prescribed.

6.4 Discussion

DSCN is introduced as a method that uses tissue and cell-line models to discovery and rank target combinations according to their estimated lethality. With existing evidence from literature and database, it is demonstrated that:

1. Using known SL pairs and non-SL pairs (random combinations of genes in these SL pairs) as inputs, DSCN clearly distinguishes two groups by their TI scores. This is mainly due to the selected routing method has the best capability of distinguishing two groups over other routing methods.

2. Predictions from DSCN directly overlaps with documented SL pairs while other methods (VIPER and Option) failed to provide any direct overlaps between their predictions and documented SL pairs. VIPER and Opticon first predicts master regulators and then estimate the synergies among them, which limit their searching space. Two master regulators are not necessarily synergistic even though the subnetworks they control contain a lot of SL pairs. In contrast, even though DSCN is designed to predict combinations among targets that are associated with FDA

approved drugs, it is capable of predict any gene pairs that have expression value and essentiality, which broaden the searching space. Any potential SL pairs within the entire genome can be potentially discovered. Moreover, predictions from DSCN directly reflect the SL between two genes, which are more direct to understand and easier to be validated.

3. Predicted ranks from DSCN align with documented SL pairs in terms of SL intensity, especially for top ranked predictions (top 20, Spearman corr=0.43). This shows great accuracy of predictions. Predicted ranks also align well with (Spearman corr=0.34) documented PDAC specific SL pairs, indicating a reasonably good specificity of the DSCN predictions under context specific scenario.

Additionally, compared to VIPER and Opticon, DSCN requires the least computational time. VIPER requires ARACNe to pre-compute mutual information network, which cost roughly 6 days to generate with a whole genome scale. ARACNe does not have option to run in parallel. Opticon distributes whole task into hundreds of jobs, each of which requires various hours (from 3 hours to 50 hours) to complete. This process needs to be performed twice: one for disease-specific networks and one for null networks. Computational work of DSCN mainly consists of two parts: spectral clustering and scoring targets in subnetworks. The first part requires 2-5 hours and second part requires 12-24 hours for examining all possible combinations of all 1,423 targets within the whole genome. The whole process can be even accelerated using multiple threads as DSCN allows user to flexibly choose the number of threads in parallel.

DSCNi undergoes the similar processes as DSCN to discover and rank target combinations for individuals. One TNBC sample from TCGA and one TNBC cell-line sample from CCLE are selected as input to predict target combinations. The predictions are validated using the following schemes:

1. The predictions are compared with existing drug combination synergies. Predicted synergistic target combinations has a high positive correlation (0.32) with target combinations reflected in synergistic drug combinations (bliss>0.12). Moreover, predictions from DSCNi doesn't have false positives, which is critical for the predictions to be applied in clinic.

2. Drug combinations associated with top ranked synergistic target combinations are either widely used in clinic as treatment plans or frequently reported as novel treatment plans in literatures. Combinations containing As_2O_3 are top ranked because their associated target combinations are in the upstream and downstream areas of Ras/MAPK pathway, which have been already targeted for successfully treating hematologic malignancies and melanomas. The findings might benefit the As_2O_3 associated combinational therapies to be re-purposed on treating TNBC.

3. For the top ranked target combinations, similarities between tissue subnetworks with and cell-line subnetworks are measured. Two networks are reasonably similar in terms of expression patterns and gene-gene correlations. This also strengthens the fact that As_2O_3 associated combinational therapies can be re-purposed.

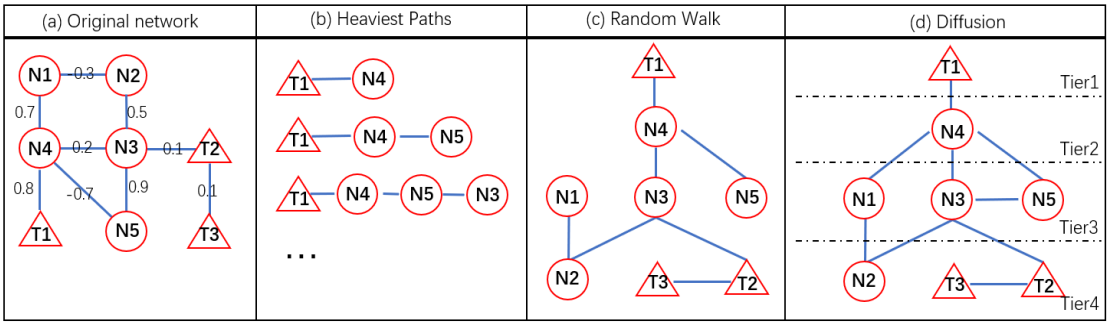

Figure 14 Three different routing methods (a) Original network. The numbers indicate edge weights (b) Heaviest paths between target node T_1 are all the other nodes (c) Routes constructed by random walk process (d) Hierarchical structures constructed by diffusion process

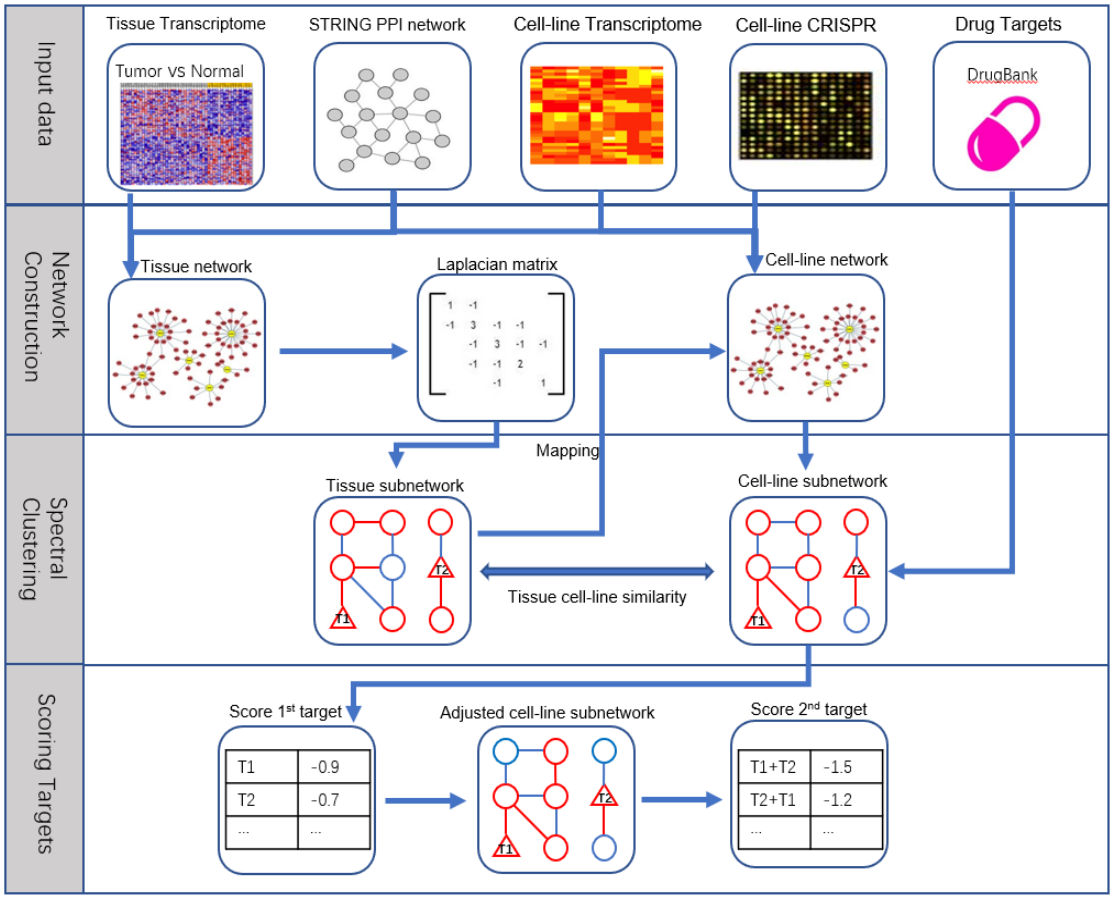

Figure 15 DSCN and DSCNi workflow

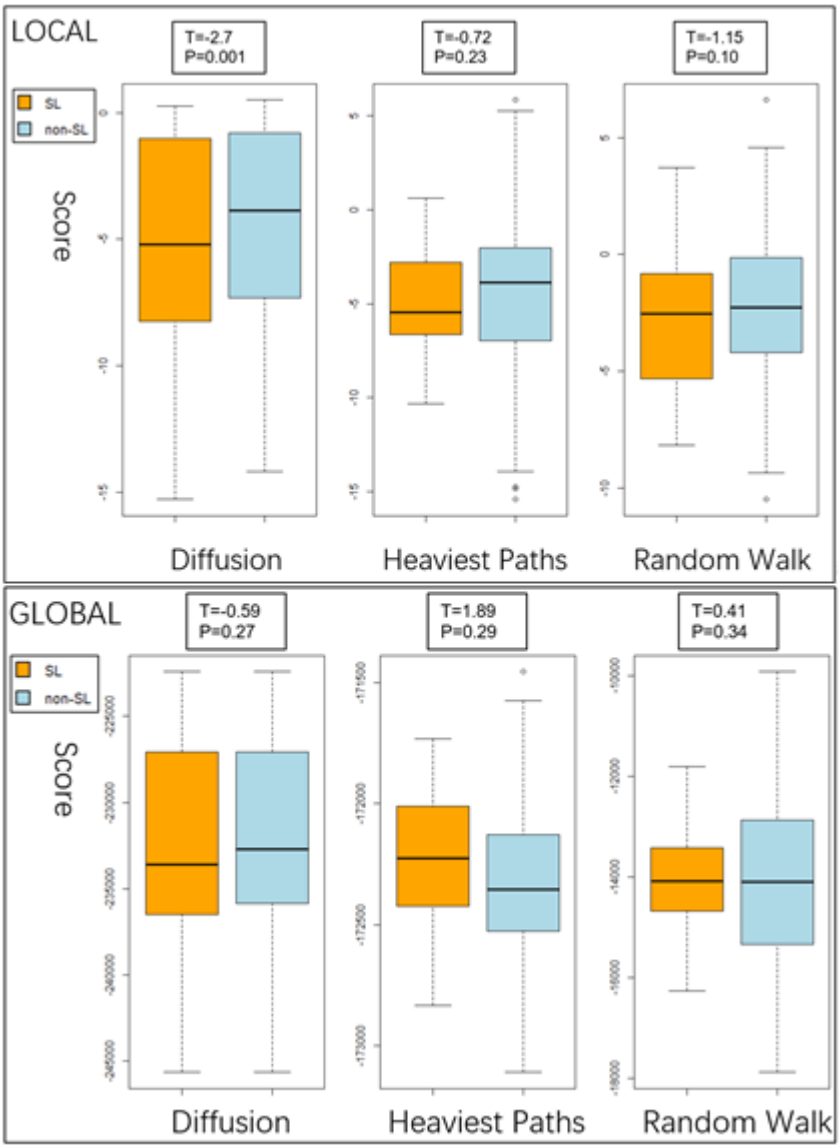

Figure 16 Score difference under different routing methods

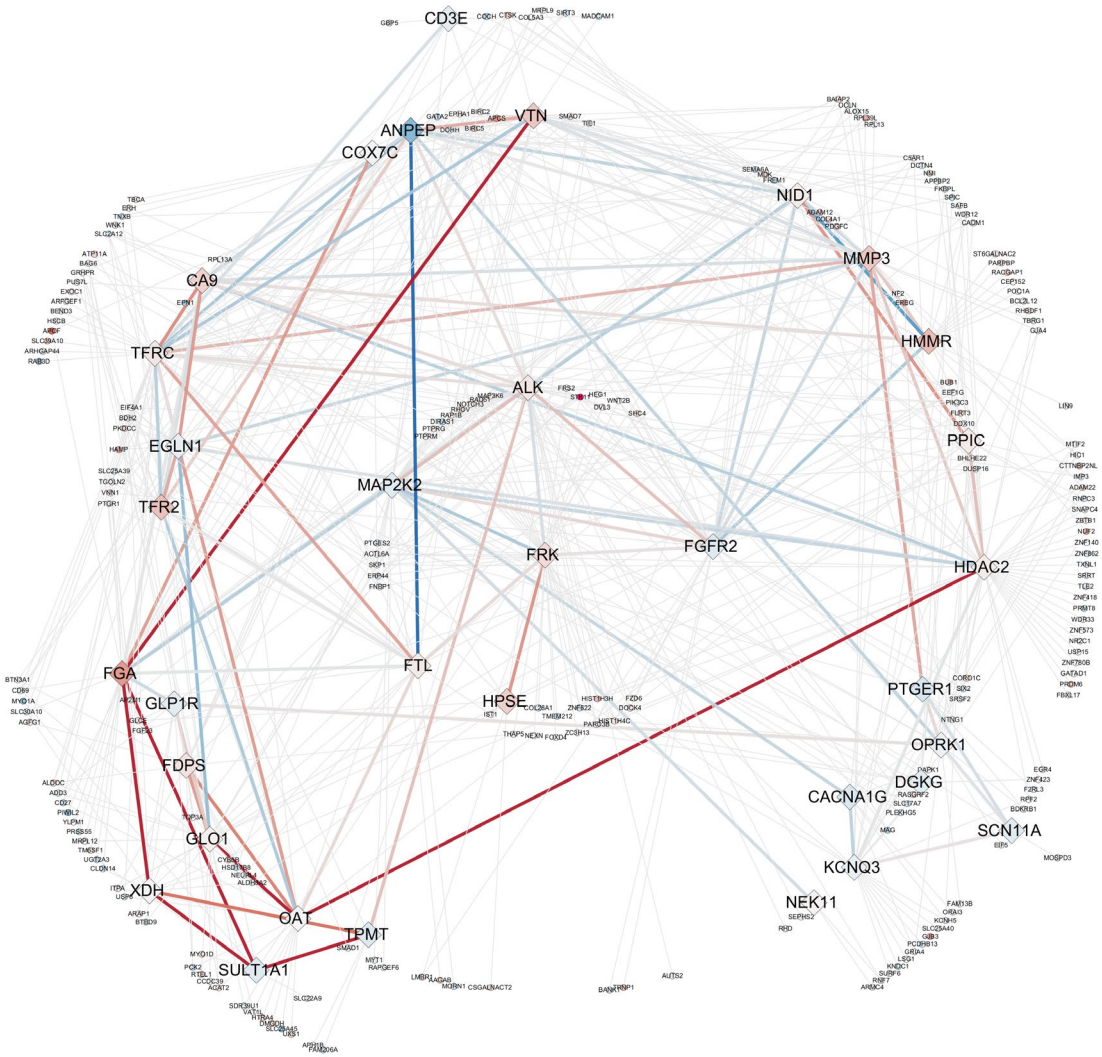

Figure 17 Subnetwork No.42 in tissue specific network

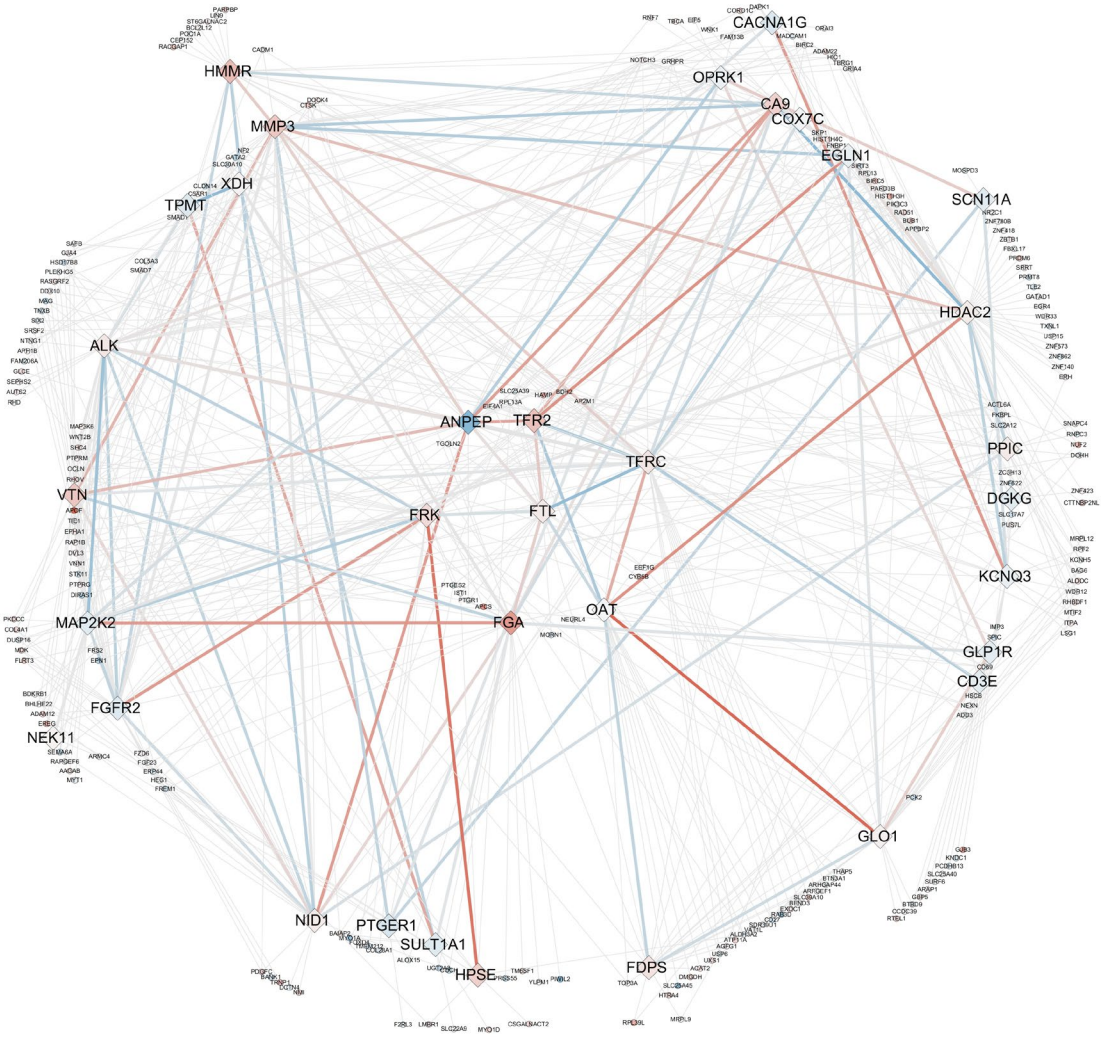

Figure 18 Subnetwork No.42 in cell-line specific network

Table 7 Compositions and sources of pancreatic omics-data

GEO Expression data			Perturbation data	PPI network	Drug target
Cancer cell-line samples	Cancer tissue samples	Normal tissue samples	CRISPR-Cas9 screening data from 'Project Achilles' (v3.3.8)	STRING PPI network for human (v11).	Targets of all FDA approved drugs from Drugbank[139]
GSE36133 (43)	GSE42952 (33)	GSE46385 (3)			
GSE46385 (7)	GSE51978 (2)	GSE16515 (16)			
GSE21654 (22)	GSE16515 (36)	GSE15471 (39)			
GSE17891 (20)	GSE15471 (39)				
	GSE23952 (3)				
92 samples	113 samples	58 samples	26 cell-lines	11,609,230 connections	1,769 targets
TCGA expression data					
	179 samples	4 samples			

Table 8 Spearman correlations between predicted target combinations and documented SL pairs

Type	Spearman Correlation	P-value	SL vs Non-SL	T-statistic	P-value
PDAC	0.34	<0.01	PDAC	0.34	<0.01
Top 20	0.43	0.06	Top 20	0.41	0.06
Top 40	0.25	0.11	Top 40	0.24	0.12
Top 100	0.27	0.02	Top 100	0.24	0.01
Total 936	0.16	<0.01	Total 936	0.15	<0.01

Table 9 Top ranked target combinations and their statistics

Gene1	Gene2	TI score	Hazard Ratio (high vs low)	P-value	Pathways in concert with Ferroptosis
EGLN1	TFRC	-255.12	2.00	0.02	Hypoxia; Ferroptosis
MAP2K2	TFRC	-255.05	1.60	0.08	MAPK; Ferroptosis
HPSE	TFRC	-255.01	1.50	0.06	Metabolism; Ferroptosis
PPIC	TFRC	-254.86	1.80	0.05	Immune system; Ferroptosis
FRK	TFRC	-254.86	1.80	0.05	Immune system; Ferroptosis
EGLN1	COX7C	-254.79	1.10	0.85	Hypoxia; Metabolism
XDH	TFRC	-254.75	2.40	<0.01	Metabolism; Ferroptosis
MAP2K2	COX7C	-254.72	0.65	0.15	MAPK; Oxidative phosphorylation
FTL	TFRC	-254.71	1.60	0.1	Ferroptosis; Ferroptosis
HPSE	COX7C	-254.68	1.10	0.7	Metabolism; Oxidative phosphorylation

Table 10 Contingency table of predicted synergy and actual drug synergy

Type	PS	PN	Correlations
DS	2594	7097	$\phi = 0.32$
DN	0	4375	

Table 11 Top ranked and selected target combinations and corresponding drug combinations from DSCNi

Gene1	Gene2	Drug1	Drug2	Bliss	TI_score	Synleth_DB score
BCL2	ERBB2	PACLITAXEL	LAPATINIB	13	-5.73	0.1775
ABCC10	RRM1	DOCETAXEL	HYDROXYUREA	3.5	-6.96	0.5
SMO	MAP4	VISMODEGIB	PACLITAXEL	14.93	-12.93	0.5
DCK	RRM1	PEMETREXED	HYDROXYUREA	10.67	-8.80	0.5
SMO	NR1I2	VISMODEGIB	PACLITAXEL	14.93	-8.76	0.5
RRM1	MAPK3	HYDROXYUREA	ARSENIC TRIOXIDE	6.63	-69.80	NA
PGR	MAPK3	MEGESTROL ACETATE	ARSENIC TRIOXIDE	10.1	-64.00	NA
SMO	MAPK3	VISMODEGIB	ARSENIC TRIOXIDE	9.03	-65.62	NA

7. Conclusion and Future Work

7.1 Conclusion on FFBN algorithm (section 3)

The benchmark studies show that GENIE3 has overall best AUROC and AUPR in silico network, and equally good AUROC but worse AUPR in yeast network compared to FFBN and FGS. Unlike GENIE3, FFBN infers the GRN purely in a data-driven way. It doesn't rely on existing prior knowledge to infer the network structure. Speed-wise, prior knowledge of TF and no-direction-inferring feature provides GENIE3 the fastest speed among the three methods. FFBN maintains a significantly faster speed under different parameters and networks with different scales compared to FGS. The speed increase varies from 11% up to 96%, and the speed difference between FFBN and FGS becomes larger as the network becomes denser or larger. Taken together, FFBN shows ascendancy over FGS when reconstructing large and dense biological networks. This computation advantage is also reflected in the GRN constructions for CLM, PC, and PL cancer samples. FFBN was able to build up these GRNs in between 2,430 and 5,931 minutes, while FGS failed to generate a converged PL GRN, prohibiting any follow-up network comparison and pathway enrichment analyses.

7.2 Conclusion on SCN algorithm (section 4)

SCN is the first algorithm that can incorporate expression data, PPI data and gene perturbation data (CRISPR or RNAi) for selecting and ranking drug targets. The novelty of SCN algorithm mainly reflects on: i. SCN is the first algorithm that takes advantage of dimension reduction methods to integrate three different types of omics

data into a comprehensive network for drug target selection; ii. SCN utilized CRISPR data to benefit the target selection. The CRISPR data can mimic the real drug response of drugs; iii. SCN uses spectral clustering to reduce data dimensions to captures features on tissue-based omics-data and ranks drug targets on cell-line omics-data, which makes the target selection process more reliable. Spectral clustering was initially introduced to cancer biology for identifying novel subtypes of Triple Negative Breast Cancer (TNBC) [155]. To our knowledge, it has never been used for selecting genotypic features from an integrated network. Despite the advantages, there is still room for SCN to improve. The possible future might include i. incorporate pathway information into target selection process for PDAC. Pathways information provide a different perspective of understanding the progression and treatment of PDAC [96, 156, 157]. Targeting cancer related pathways can be a highly effective strategy for treating PDAC. Thus, it is necessary to incorporate pathway information into the drug target ranking and selection process; ii. Incorporate functional information into the target selection process. SCN algorithm ranked drug targets mainly based on differential expression, protein-protein interaction and tissue-target concordance. However, different proteins might have different docking capacities, which directly affects its potential to become a druggable target. Unfortunately, SCN algorithm does not take this information into account for ranking targets. Integrating this information into the whole process is necessary.

7.3 Conclusion on SCN website (section 5)

SCN website offers a unique method of prioritizing actionable drug targets from multi-type omics data, including tissue and cell-line expression profiles, PPIs, CRISPR screening data and drug target information from DrugBank. Over 20,000 genes, 9606 proteins, expression profile from >400 cell-lines across 24 different cancer types are included. For the first time in precision medicine, this platform integrates tissue data and cell-line data of cancers, enabling users to upload their own expression data to seek customized targets. Survival analysis is automatically performed using TCGA patient data so that users can validate their targets. Moreover, SCN website provides GSEA analysis for the prioritized targets to better mine the biological mechanisms associated with them. Additionally, evidence of clinical trials extracted from ClincalTrials.gov can provide more hints on how the identified targets can be applied in clinic.

SCN website provides instructions of input, output files and general steps of SCN algorithm. It offers an complete set of example files, which contains example tissue expression data, cell-line expression data, all intermediate files generated during the algorithm processing and example output file so that users can easily understand and try the whole process.

7.4 Conclusion on DSCN algorithm (section 6)

DSCN and DSCNi has been validated as solid approaches for predicting lethality for combinational targets. Some of these combinational targets either serve as widely used in clinic. Other top ranked combinational targets are extensively studied

as promising treatment plans. Taken together, these predictions worth further investigation for either investigating the molecular mechanisms of cancer or developing novel treatment plans.

References

1. Vidal, M.J.F.I., *A unifying view of 21st century systems biology.* 2009. 583(24): p. 3891-3894.
2. Peters, A., *A combination of cis and trans control can solve the hotspot conversion paradox.* Genetics, 2008. 178(3): p. 1579-1593.
3. Dong, X. and Z. Weng, *The correlation between histone modifications and gene expression.* Epigenomics, 2013. 5(2): p. 113-116.
4. Phillips, T., *The role of methylation in gene expression.* Nature Education, 2008. 1(1): p. 116.
5. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. 28(1): p. 27-30.
6. Cerami, E.G., et al., *Pathway Commons, a web resource for biological pathway data.* Nucleic Acids Res, 2011. 39(Database issue): p. D685-90.
7. Karp, P.D., et al., *The metacyc database.* Nucleic acids research, 2002. 30(1): p. 59-61.
8. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic acids research, 2007. 36(suppl_1): p. D102-D106.
9. Draizen, E.J., et al., *HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants.* Database, 2016. 2016.
10. Cho, S., et al., *MiRGator v3. 0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting.* Nucleic acids research, 2012. 41(D1): p. D252-D257.
11. Fishilevich, S., et al., *GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.* Database, 2017. 2017.
12. Cheng, C., et al., *Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.* PLoS computational biology, 2011. 7(11).
13. Tong, A.H.Y., et al., *Global mapping of the yeast genetic interaction network.* 2004. 303(5659): p. 808-813.
14. Margolin, A.A., et al. *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*. in *BMC bioinformatics*. 2006. BioMed Central.
15. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* 2005. 122(6): p. 957-968.
16. Venkatesan, K., et al., *An empirical framework for binary interactome mapping.* Nature methods, 2009. 6(1): p. 83.
17. Baryshnikova, A., et al., *Genetic interaction networks: toward an understanding of heritability.* Annual review of genomics and human genetics, 2013. 14: p. 111-133.

18. Madan Babu, M. and S.A. Teichmann, *Evolution of transcription factors and the gene regulatory network in Escherichia coli.* Nucleic acids research, 2003. 31(4): p. 1234-1244.
19. Szklarczyk, D., et al., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.* Nucleic Acids Res, 2017. 45(D1): p. D362-D368.
20. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nature methods, 2014. 11(3): p. 333.
21. Costanzo, M., et al., *A global genetic interaction network maps a wiring diagram of cellular function.* Science, 2016. 353(6306): p. aaf1420.
22. Costanzo, M., et al., *The genetic landscape of a cell.* science, 2010. 327(5964): p. 425-431.
23. Hart, T., et al., *High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities.* Cell, 2015. 163(6): p. 1515-1526.
24. Wang, T., et al., *Identification and characterization of essential genes in the human genome.* Science, 2015. 350(6264): p. 1096-1101.
25. Blomen, V.A., et al., *Gene essentiality and synthetic lethality in haploid human cells.* Science, 2015. 350(6264): p. 1092-1096.
26. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* science, 2003. 302(5643): p. 249-255.
27. Barabási, A.-L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nature reviews genetics, 2011. 12(1): p. 56.
28. Elo, L.L., et al., *Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process.* Bioinformatics, 2007. 23(16): p. 2096-2103.
29. Luo, F., et al., *Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory.* BMC bioinformatics, 2007. 8(1): p. 299.
30. Terentiev, A., N. Moldogazieva, and K. Shaitan, *Dynamic proteomics in modeling of the living cell. Protein-protein interactions.* Biochemistry (Moscow), 2009. 74(13): p. 1586-1607.
31. Brettner, L.M. and J. Masel, *Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast.* BMC systems biology, 2012. 6(1): p. 128.
32. Szklarczyk, D., et al., *The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible.* Nucleic acids research, 2016: p. gkw937.
33. O'brien, K.P., M. Remm, and E.L. Sonnhammer, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic acids research, 2005. 33(suppl_1): p. D476-D480.
34. Zhang, Q.C., et al., *Structure-based prediction of protein–protein interactions on a genome-wide scale.* Nature, 2012. 490(7421): p. 556-560.

35. Schmitt, T., C. Ogris, and E.L. Sonnhammer, *FunCoup 3.0: database of genome-wide functional coupling networks.* Nucleic acids research, 2014. 42(D1): p. D380-D388.
36. Vinayagam, A., et al., *A directed protein interaction network for investigating intracellular signal transduction.* Science signaling, 2011. 4(189): p. rs8-rs8.
37. Clarke, B., *Information optimality and Bayesian modelling.* Journal of Econometrics, 2007. 138(2): p. 405-429.
38. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* Journal of computational biology, 2000. 7(3-4): p. 601-620.
39. Mukherjee, S. and T.P. Speed, *Network inference using informative priors.* Proceedings of the National Academy of Sciences, 2008. 105(38): p. 14313-14318.
40. Chickering, D.M., *Optimal structure identification with greedy search.* Journal of machine learning research, 2002. 3(Nov): p. 507-554.
41. Hill, S.M., et al., *Inferring causal molecular networks: empirical assessment through a community-based effort.* Nature methods, 2016. 13(4): p. 310-318.
42. Pearl, J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*. 2014: Elsevier.
43. Cho, D.-Y., Y.-A. Kim, and T.M. Przytycka, *Network biology approach to complex diseases.* PLoS computational biology, 2012. 8(12).
44. Chen, J. and B. Yuan, *Detecting functional modules in the yeast protein–protein interaction network.* Bioinformatics, 2006. 22(18): p. 2283-2290.
45. Feng, J., R. Jiang, and T. Jiang, *A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010. 8(3): p. 621-634.
46. Maraziotis, I.A., K. Dimitrakopoulou, and A. Bezerianos, *Growing functional modules from a seed protein via integration of protein interaction and gene expression data.* Bmc Bioinformatics, 2007. 8(1): p. 408.
47. Huang, L., et al., *Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction.* Bioinformatics, 2019. 35(19): p. 3709-3717.
48. Huang, L., et al., *DrugComboRanker: drug combination discovery based on target network analysis.* Bioinformatics, 2014. 30(12): p. i228-i236.
49. Hu, Y., et al., *Optimal control nodes in disease-perturbed networks as targets for combination therapy.* Nature communications, 2019. 10(1): p. 1-14.
50. Cohen, R., et al., *Resilience of the internet to random breakdowns.* Physical review letters, 2000. 85(21): p. 4626.
51. Cai, C., et al., *Is human blood a good surrogate for brain tissue in transcriptional studies?* BMC genomics, 2010. 11(1): p. 589.
52. Liu, Z., B. Malone, and C. Yuan. *Empirical evaluation of scoring functions for Bayesian network model selection*. in *BMC bioinformatics*. 2012. Springer.
53. Consortium, G.O., *The Gene Ontology in 2010: extensions and refinements.* Nucleic acids research, 2010. 38(suppl_1): p. D331-D335.

54. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic acids research, 2006. 34(suppl_1): p. D535-D539.
55. Marbach, D., et al., *Wisdom of crowds for robust gene network inference.* Nature methods, 2012. 9(8): p. 796.
56. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. 313(5795): p. 1929-35.
57. Ma, Y., et al., *Prediction of candidate drugs for treating pancreatic cancer by using a combined approach.* PloS one, 2016. 11(2).
58. Alvarez, M.J., et al., *Functional characterization of somatic mutations in cancer using network-based inference of protein activity.* Nature genetics, 2016. 48(8): p. 838.
59. Ramsey, J.D., *Scaling up greedy causal search for continuous variables.* arXiv preprint arXiv:1507.07749, 2015.
60. Dennis, G., et al., *DAVID: database for annotation, visualization, and integrated discovery.* Genome biology, 2003. 4(9): p. R60.
61. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic acids research, 2002. 30(1): p. 207-210.
62. Meek, C., *Causal inference and causal explanation with background knowledge.* arXiv preprint arXiv:1302.4972, 2013.
63. Irrthum, A., L. Wehenkel, and P. Geurts, *Inferring regulatory networks from expression data using tree-based methods.* PloS one, 2010. 5(9): p. e12776.
64. Chakravarty, D., et al., *OncoKB: a precision oncology knowledge base.* JCO precision oncology, 2017. 1: p. 1-16.
65. Voutsadakis, I.A., *The ubiquitin–proteasome system and signal transduction pathways regulating epithelial mesenchymal transition of cancer.* Journal of biomedical science, 2012. 19(1): p. 67.
66. Giovannucci, E., *Insulin and colon cancer.* Cancer Causes & Control, 1995. 6(2): p. 164-179.
67. Mihaylova, M.M. and R.J. Shaw, *The AMPK signalling pathway coordinates cell growth, autophagy and metabolism.* Nature cell biology, 2011. 13(9): p. 1016.
68. Li, W., et al., *Targeting AMPK for cancer prevention and treatment.* Oncotarget, 2015. 6(10): p. 7365.
69. Vander Heiden, M.G., L.C. Cantley, and C.B. Thompson, *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* science, 2009. 324(5930): p. 1029-1033.
70. Ashton, T.M., et al., *Oxidative phosphorylation as an emerging target in cancer therapy.* Clinical Cancer Research, 2018. 24(11): p. 2482-2490.
71. Solaini, G., G. Sgarbi, and A. Baracca, *Oxidative phosphorylation in cancer cells.* Biochimica et Biophysica Acta (BBA)-Bioenergetics, 2011. 1807(6): p. 534-542.

72. Gründker, C. and G. Emons, *The role of gonadotropin-releasing hormone in cancer cell proliferation and metastasis.* Frontiers in endocrinology, 2017. 8: p. 187.
73. Luo, Y., et al., *A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information.* Nat Commun, 2017. 8(1): p. 573.
74. Dimitrakopoulos, C., et al., *Network-based integration of multi-omics data for prioritizing cancer genes.* Bioinformatics, 2018. 34(14): p. 2441-2448.
75. Ritchie, M.D., et al., *Methods of integrating data to uncover genotype-phenotype interactions.* Nat Rev Genet, 2015. 16(2): p. 85-97.
76. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. 28(1): p. 27-30.
77. Nielsen, T.E. and S.L. Schreiber, *Towards the optimal screening collection: a synthesis strategy.* Angew Chem Int Ed Engl, 2008. 47(1): p. 48-56.
78. Wang, S. and J. Peng, *Network-assisted target identification for haploinsufficiency and homozygous profiling screens.* PLoS Comput Biol, 2017. 13(6): p. e1005553.
79. Ferrero, E., I. Dunham, and P. Sanseau, *In silico prediction of novel therapeutic targets using gene–disease association data.* Journal of translational medicine, 2017. 15(1): p. 182.
80. Tsherniak, A., et al., *Defining a Cancer Dependency Map.* Cell, 2017. 170(3): p. 564-576 e16.
81. Aguirre, A.J., et al., *Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting.* Cancer Discov, 2016. 6(8): p. 914-29.
82. Cowley, G.S., et al., *Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies.* Sci Data, 2014. 1: p. 140035.
83. Lin, A., et al., *CRISPR/Cas9 mutagenesis invalidates a putative cancer dependency targeted in on-going clinical trials.* Elife, 2017. 6: p. e24179.
84. Wei, Y.-C. and C.-K. Cheng. *Towards efficient hierarchical designs by ratio cut partitioning*. in *1989 IEEE International Conference on Computer-Aided Design*. 1989. IEEE.
85. Shi, J. and J. Malik, *Normalized cuts and image segmentation.* IEEE Transactions on pattern analysis and machine intelligence, 2000. 22(8): p. 888-905.
86. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. 28(1): p. 100-108.
87. Chiang, M.M.-T. and B. Mirkin, *Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads.* Journal of classification, 2010. 27(1): p. 3-40.
88. Wang, J., et al., *A glycolytic mechanism regulating an angiogenic switch in prostate cancer.* Cancer Res, 2007. 67(1): p. 149-59.

89. Zieker, D., et al., *Phosphoglycerate kinase 1 a promoting enzyme for peritoneal dissemination in gastric cancer.* Int J Cancer, 2010. 126(6): p. 1513-20.
90. Wang, J., et al., *Characterization of phosphoglycerate kinase-1 expression of stromal cells derived from tumor microenvironment in prostate cancer progression.* Cancer Res, 2010. 70(2): p. 471-80.
91. Punjabi, P. and A. Murday, *Successful surgical repair of a false aneurysm of the ascending aorta following orthotopic cardiac transplantation: a case report.* Eur J Cardiothorac Surg, 1997. 11(6): p. 1174-5.
92. Kalmyrzaev, B., et al., *Hyaluronan-mediated motility receptor gene single nucleotide polymorphisms and risk of breast cancer.* Cancer Epidemiol Biomarkers Prev, 2008. 17(12): p. 3618-20.
93. Shigeishi, H., et al., *Overexpression of the receptor for hyaluronan-mediated motility, correlates with expression of microtubule-associated protein in human oral squamous cell carcinomas.* Int J Oncol, 2009. 34(6): p. 1565-71.
94. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. 34(3): p. 267-73.
95. Tang, Z., et al., *GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses.* Nucleic Acids Res, 2017. 45(W1): p. W98-W102.
96. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. 45(10): p. 1113-20.
97. Carithers, L.J., et al., *A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project.* Biopreserv Biobank, 2015. 13(5): p. 311-9.
98. Amanam, I. and V. Chung, *Targeted Therapies for Pancreatic Cancer.* Cancers (Basel), 2018. 10(2).
99. Yamanaka, Y., et al., *Overexpression of HER2/neu oncogene in human pancreatic carcinoma.* Hum Pathol, 1993. 24(10): p. 1127-34.
100. Chou, A., et al., *Clinical and molecular characterization of HER2 amplified-pancreatic cancer.* Genome Med, 2013. 5(8): p. 78.
101. Li, X., et al., *Mitochondria-Translocated PGK1 Functions as a Protein Kinase to Coordinate Glycolysis and the TCA Cycle in Tumorigenesis.* Mol Cell, 2016. 61(5): p. 705-719.
102. Hu, H., et al., *Acetylation of PGK1 promotes liver cancer cell proliferation and tumorigenesis.* Hepatology, 2017. 65(2): p. 515-528.
103. Xie, H., et al., *PGK1 Drives Hepatocellular Carcinoma Metastasis by Enhancing Metabolic Process.* Int J Mol Sci, 2017. 18(8).
104. Rajeshkumar, N.V., et al., *Therapeutic Targeting of the Warburg Effect in Pancreatic Cancer Relies on an Absence of p53 Function.* Cancer Res, 2015. 75(16): p. 3355-64.
105. Grutzmann, R., et al., *Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays.* Neoplasia, 2004. 6(5): p. 611-22.

106. Tzankov, A., et al., *In situ RHAMM protein expression in acute myeloid leukemia blasts suggests poor overall survival.* Ann Hematol, 2011. 90(8): p. 901-9.
107. Yamano, Y., et al., *Hyaluronan-mediated motility: a target in oral squamous cell carcinoma.* Int J Oncol, 2008. 32(5): p. 1001-9.
108. Ishigami, S., et al., *Prognostic impact of CD168 expression in gastric cancer.* BMC Cancer, 2011. 11: p. 106.
109. Du, Y.C., et al., *Receptor for hyaluronan-mediated motility isoform B promotes liver metastasis in a mouse model of multistep tumorigenesis and a tail vein assay for metastasis.* Proc Natl Acad Sci U S A, 2011. 108(40): p. 16753-8.
110. Maxwell, C.A., et al., *Interplay between BRCA1 and RHAMM regulates epithelial apicobasal polarization and may influence risk of breast cancer.* PLoS biology, 2011. 9(11): p. e1001199.
111. Amano, T., et al., *Antitumor effects of vaccination with dendritic cells transfected with modified receptor for hyaluronan-mediated motility mRNA in a mouse glioma model.* Journal of neurosurgery, 2007. 106(4): p. 638-645.
112. Willemen, Y., et al., *The tumor-associated antigen RHAMM (HMMR/CD168) is expressed by monocyte-derived dendritic cells and presented to T cells.* Oncotarget, 2016. 7(45): p. 73960-73970.
113. Li, J., X. Ji, and H. Wang, *Targeting Long Noncoding RNA HMMR-AS1 Suppresses and Radiosensitizes Glioblastoma.* Neoplasia, 2018. 20(5): p. 456-466.
114. Li, J., et al., *Knockdown of POLE2 expression suppresses lung adenocarcinoma cell malignant phenotypes in vitro.* Oncol Rep, 2018. 40(5): p. 2477-2486.
115. Li, W., et al., *MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens.* Genome biology, 2014. 15(12): p. 554.
116. Dimitrakopoulos, C., et al., *Network-based integration of multi-omics data for prioritizing cancer genes.* Bioinformatics, 2018. 34(14): p. 2441-2448.
117. Rauscher, B., et al., *GenomeCRISPR-a database for high-throughput CRISPR/Cas9 screens.* Nucleic acids research, 2016: p. gkw997.
118. Tsherniak, A., et al., *Defining a cancer dependency map.* Cell, 2017. 170(3): p. 564-576. e16.
119. Behan, F.M., et al., *Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens.* Nature, 2019. 568(7753): p. 511.
120. Aguirre, A., et al., *Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. Cancer Discov. 2016; 6: 914–929. doi: 10.1158/2159-8290*. CD-16-0154.[PMC free article][PubMed][Cross Ref].
121. Cowley, G.S., et al., *Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies.* Scientific data, 2014. 1: p. 140035.
122. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic acids research, 2013. 42(D1): p. D1091-D1097.

123. Tang, Z., et al., *GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses.* Nucleic acids research, 2017. 45(W1): p. W98-W102.
124. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019.* CA: a cancer journal for clinicians, 2019. 69(1): p. 7-34.
125. Iovanna, J., et al., *Current knowledge on pancreatic cancer.* Front Oncol, 2012. 2: p. 6.
126. Kruger, S., et al., *Translational research in pancreatic ductal adenocarcinoma: current evidence and future concepts.* World J Gastroenterol, 2014. 20(31): p. 10769-77.
127. Kamisawa, T., et al., *Pancreatic cancer.* The Lancet, 2016. 388(10039): p. 73-85.
128. Domenichini, A., et al., *Pancreatic cancer tumorspheres are cancer stem-like cells with increased chemoresistance and reduced metabolic potential.* Adv Biol Regul, 2019. 72: p. 63-77.
129. Parhi, P., C. Mohanty, and S.K. Sahoo, *Nanotechnology-based combinational drug delivery: an emerging approach for cancer therapy.* Drug discovery today, 2012. 17(17-18): p. 1044-1052.
130. Gillet, J.-P., S. Varma, and M.M. Gottesman, *The clinical relevance of cancer cell lines.* Journal of the National Cancer Institute, 2013. 105(7): p. 452-458.
131. Frese, K.K. and D.A. Tuveson, *Maximizing mouse cancer models.* Nature Reviews Cancer, 2007. 7(9): p. 654.
132. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nature protocols, 2013. 8(11): p. 2281-2308.
133. Shi, J., et al., *Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains.* Nature biotechnology, 2015. 33(6): p. 661.
134. Wang, T., E.S. Lander, and D.M. Sabatini, *Large-Scale Single Guide RNA Library Construction and Use for CRISPR-Cas9-Based Genetic Screens.* Cold Spring Harb Protoc, 2016. 2016(3): p. pdb top086892.
135. Vincent, A., et al., *Pancreatic cancer.* The lancet, 2011. 378(9791): p. 607-620.
136. Liu, Y.-Y., J.-J. Slotine, and A.-L. Barabási, *Controllability of complex networks.* nature, 2011. 473(7346): p. 167-173.
137. Sun, Y., et al., *Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer.* Nature communications, 2015. 6(1): p. 1-10.
138. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. 483(7391): p. 603-607.
139. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic acids research, 2006. 34(suppl_1): p. D668-D672.
140. Guo, J., H. Liu, and J. Zheng, *SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets.* Nucleic acids research, 2016. 44(D1): p. D1011-D1017.

141. Shoemaker, R.H., *The NCI60 human tumour cell line anticancer drug screen.* Nature Reviews Cancer, 2006. 6(10): p. 813-823.
142. Margolin, A.A., et al. *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*. in *BMC bioinformatics*. 2006. Springer.
143. Jeong, S.M., S. Hwang, and R.H. Seong, *Transferrin receptor regulates pancreatic cancer growth by modulating mitochondrial respiration and ROS generation.* Biochemical and biophysical research communications, 2016. **471**(3): p. 373-379.
144. Xie, Y., et al., *Ferroptosis: process and function.* Cell Death & Differentiation, 2016. **23**(3): p. 369-379.
145. Borisy, A.A., et al., *Systematic discovery of multicomponent therapeutics.* Proceedings of the National Academy of Sciences, 2003. **100**(13): p. 7977-7982.
146. O'Neil, J., et al., *An unbiased oncology compound screen to identify novel combination strategies.* Molecular cancer therapeutics, 2016. **15**(6): p. 1155-1162.
147. Giltnane, J.M. and J.M. Balko, *Rationale for targeting the Ras/MAPK pathway in triple-negative breast cancer.* Discovery medicine, 2014. **17**(95): p. 275-283.
148. Jang, S. and M. Atkins, *Treatment of BRAF-mutant melanoma: the role of vemurafenib and other therapies.* Clinical Pharmacology & Therapeutics, 2014. **95**(1): p. 24-31.
149. Bartholomeusz, C., et al., *MEK inhibitor selumetinib (AZD6244; ARRY-142886) prevents lung metastasis in a triple-negative breast cancer xenograft model.* Molecular cancer therapeutics, 2015. **14**(12): p. 2773-2781.
150. Nagaria, T.S., et al., *Combined targeting of Raf and Mek synergistically inhibits tumorigenesis in triple negative breast cancer model systems.* Oncotarget, 2017. **8**(46): p. 80804.
151. Xin, X., et al., *Inhibition of FEN1 Increases Arsenic Trioxide-Induced ROS Accumulation and Cell Death: Novel Therapeutic Potential for Triple Negative Breast Cancer.* Frontiers in Oncology, 2020. **10**.
152. Alipour, F., et al., *Inhibition of PI3K pathway using BKM120 intensified the chemo-sensitivity of breast cancer cells to arsenic trioxide (ATO).* The international journal of biochemistry & cell biology, 2019. **116**: p. 105615.
153. Pei, X., et al., *Effect of tetrandrine combined with arsenic trioxide on stem cells of triple negative breast cancer.* 2019.
154. Miodragović, Đ., et al., *Beyond cisplatin: Combination therapy with arsenic trioxide.* Inorganica Chimica Acta, 2019. **496**: p. 119030.
155. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nat Methods, 2014. **11**(3): p. 333-7.
156. Eser, S., et al., *Oncogenic KRAS signalling in pancreatic cancer.* British journal of cancer, 2014. **111**(5): p. 817.
157. Neuzillet, C., et al., *Targeting the TGFβ pathway for cancer therapy.* Pharmacology & therapeutics, 2015. **147**: p. 22-31.

Curriculum Vitae

Enze Liu

Education

Ph.D. Bioinformatics —Indiana University-Perdue University Indianapolis, Indianapolis, IN, USA, 2015-2020

M.S. Computational Biology (Bioinformatics track) —KTH Royal Institute of Technology, Sweden 2010-2013

B.E. Computer Science — Jiangsu Normal University, China 2005-2009

Work experience

Research Assistant—Indiana University, Indianapolis, IN, USA, 2015-2020

Bioinformatics Engineer—Biomarker Bio-technology corporation, Beijing, China 2014-2014

Bioinformatics researcher—Department of Biophysics and Biochemistry, Stockholm University; Science for life laboratory, Karolinska Institutet Science Park; Sweden 2013-2013

Publications

Liu E, Li, L. D-SCN: A Spectral clustering approach for prioritizing target combinations for pancreatic ductal adenocarcinoma. (In preparation)

Liu E, Poffenberger L, Garge P et al. SCN: a website for prioritizing cancer therapeutic targets using CRISPR-Cas9 data and network model[J]. Journal of Nucleic Acid Research, 2020 (Under review)

Liu E, Zhang Z, Cheng X et al. SCN: Spectral Clustering for Network-based Ranking to reveal potential drug targets and its application in pancreatic ductal adenocarcinoma[J]. BMC Medical Genomics, 2020.

Liu E, Kinnebrew G, Li J, et al. A Fast and Furious Bayesian Network and Its Application of Identifying Colon Cancer to Liver Metastasis Gene Regulatory Networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019.

Liu E, Li L, Cheng L. Gene Regulatory Network Review[J]. 2019.

Liu E. Optimization and Application extension fora Bloom filter-based sequence classifier[J]. 2013.

Cheng L, Pandya P H, Liu E, et al. Integration of genomic copy number variations and chemotherapy-response biomarkers in pediatric sarcoma[J]. BMC medical genomics, 2019, 12(1): 23. Zhang J, Zhang Q, Cheng T, et al.

Wang L, Li J, Liu E, et al. Identification of Alternatively Activated Pathways between Primary Breast Cancer and Liver Metastatic Cancer Using Microarray Data[J]. Genes, 2019, 10(10): 753.

Vasudevaraja V, Renbarger J, Shah R G, et al. PMTDS: a computational method based on genetic interaction networks for precision medicine target-drug selection in cancer[J]. Quantitative Biology, 2017, 5(4): 380-394.

Zhang J, Zhang Q, Cheng T, et al. High-density genetic map construction and identification of a locus controlling weeping trait in an ornamental woody plant (Prunus mume Sieb. et Zucc) [J]. DNA research, 2015, 22(3): 183-191.

Li J, Huo Y, Wu X, et al. Essentiality and Transcriptome-Enriched Pathway Scores Predict Drug-Combination Synergy[J]. Biology, 2020, 9(9): 278.